University of Alberta

Molecular Modelling of Protein-Protein/Protein-Solvent Interactions

by

Tyler Luchko

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Physics

Edmonton, Alberta

Fall 2008

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-46369-7
Our file Notre référence
ISBN: 978-0-494-46369-7

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Abstract

The inner workings of individual cells are based on intricate networks of protein-protein interactions. However, each of these individual protein interactions requires a complex physical interaction between proteins and their aqueous environment at the atomic scale. In this thesis, molecular dynamics simulations are used in three theoretical studies to gain insight at the atomic scale about protein hydration, protein structure and tubulin-tubulin (protein-protein) interactions, as found in microtubules. Also presented, in a fourth project, is a molecular model of solvation coupled with the Amber molecular modelling package, to facilitate further studies without the need of explicitly modelled water.

Basic properties of a minimally solvated protein were calculated through an extended study of myoglobin hydration with explicit solvent, directly investigating water and protein polarization. Results indicate a close correlation between polarization of both water and protein and the onset of protein function.

The methodology of explicit solvent molecular dynamics was further used to study tubulin and microtubules. Extensive conformational sampling of the carboxy-terminal tails of β-tubulin was performed via replica exchange molecular dynamics, allowing the characterisation of the flexibility, secondary structure and binding domains of the C-terminal tails through statistical analysis methods. Mechanical properties of tubulin and microtubules were calculated with adaptive biasing force molecular dynamics. The function of the M-loop in microtubule stability was demonstrated in these simulations. The flexibility of this loop allowed constant contacts between the protofilaments to be maintained during simulations while the smooth deformation provided a spring-like restoring force. Additionally, calculating the free energy profile between the straight and bent tubulin configurations was used to test the proposed conformational change in tubulin, thought to cause microtubule destabilization. No conformational change was observed but a nucleotide dependent 'softening' of the interaction was found instead, suggesting that an entropic force in a microtubule configuration could be the mechanism of microtubule

collapse.

Finally, to overcome much of the computational costs associated with explicit solvent calculations, a new combination of molecular dynamics with the 3D-reference interaction site model (3D-RISM) of solvation was integrated into the Amber molecular dynamics package. Our implementation of 3D-RISM shows excellent agreement with explicit solvent free energy calculations. Several optimisation techniques, including a new multiple time step method, provide a nearly 100 fold performance increase, giving similar computational performance to explicit solvent.

Acknowledgements

There are many people I need to thank and many more I should thank but I do not have much time to write this. First, I will thank my wife, Alby Luchko. She agreed to marry me in the middle of all this, gave up a much shorter last name (though I didn't ask her to), has given me day-to-day support and has read many research proposal drafts. Next I must thank my parents, who each gave me half of their DNA and then helped me to use it. Their support and Sunday dinners have been an important part of the last seven years. Also important have been the many hours of running, skiing, hockey, football and nerdy conversations I have shared with my brother, Aaron. We also share 50% of our DNA. My sisters, Leisa and Michele, and nephews, Liam and Jacob, have provided hours of sports entertainment, academic competition and money laundering/investment. Alby's family has also contributed to my success; Caroline, Charlie, Winston, Jacquie, Nathan, Jacob and Sydney have all been generous with their support and their food. Finally, I need to thank all of the Luchkos and Buttons (even if that isn't your last name).

At the University of Alberta, many people have contributed to my success. My supervisors, Jack Tuszynski and Andriy Kovalenko, have provided unwavering support and guidance. Torin Huzil, Sergey Gusarov and Piotr Drabik have all been essential in my development as a scientist. Several office mates have been critical to maintaining my sanity, but three in particular stand out: Robert Bryce, J. P. Archambault and Danielle Pahud. Of course, I need to thank all the members of the Feynmen flag football franchise, Campus Recreation 2003 Division 3 Champions and 2007 Division 1 Champions, for kick-ass football.

My computational work has been difficult and long and has required both financial and computational support. The Natural Sciences and Engineering Research Council (NSERC), National Research Council (NRC), Province of Alberta, Mathematics of Information Technology and Complex Systems (MITACS) and University of Alberta have all contributed to keeping me fed, clothed and sheltered. Westgrid, the Center for Excellence in Integrated Nanotools (CEIN) and IBM have all provided valuable computer time and expertise. Whether they know it or not, none of this would be possible without the combined efforts of Sir Tim Berners Lee, Steve Jobs, Linus Torvalds and Google.

Table of Contents

Acknowledgements

1 Introduction **1**

2 Proteins **3**

2.1 Amino Acids 3

2.2 Structure 4

2.2.1 Primary Structure 4

2.2.2 Secondary Structure 5

2.2.2.1 α Helix 5

2.2.2.2 β Sheet 6

2.2.2.3 Turns and Random Coil 6

2.2.3 Tertiary Structure 7

2.2.4 Quaternary Structure 7

2.3 Function and Role 9

2.3.1 Interaction with the Environment: Solvation 9

2.3.2 Protein-Protein Interactions 9

2.3.3 Conformational Flexibility 9

3 Microtubule Structure and Function **10**

3.1 Tubulin 10

3.1.1 MAP Binding 11

3.1.2 Lateral Contacts 11

3.1.3 Nucleotides 12
3.2 Microtubules 12
3.2.1 Structure 14
3.2.2 Dynamics 14
3.2.3 Function 17
3.2.3.1 Intracellular Transportation 17
3.2.3.2 Cellular Transportation 18
3.2.3.3 Cell Division 18
3.3 Carboxy-Terminal Tails 20
3.3.1 Physical Properties of CTTs 20
3.3.1.1 Isotypes 20
3.3.1.2 Post-translational Modifications 20
3.3.2 Biological Role 22
3.3.2.1 Subtilisin 22
3.3.2.2 Tubulin Polymerization 23
3.3.2.3 MAP Binding 23
3.3.2.4 Drug Interaction 24

4 Protein Hydration **25**
4.1 Water 25
4.1.1 Macroscopic Properties 25
4.1.2 Molecular Properties 26
4.1.2.1 Intramolecular Structure 27
4.1.2.2 Dipole Moment 27
4.1.2.3 Polarization 27
4.1.2.4 Hydrogen Bonding 28
4.1.2.5 Tetrahedral Structure 28
4.1.2.6 Intermolecular Structure 29
4.1.2.7 Proton Conduction 29
4.2 Proteins-Water Interactions 29

4.2.1 Protein Folding and Structure 29
4.2.1.1 Hydrophobicity 30
4.2.1.2 Hydrogen Bonding 31
4.2.1.3 Protein Induced Water Structure 31
4.2.2 Protein Interactions 32
4.3 Molecular Modelling of Solvent 32
4.3.1 Explicit Water Models 33
4.3.1.1 TIP3P and Family 33
4.3.1.2 SPC and Family 35
4.3.2 Implicit Solvents 35
4.3.2.1 Poisson-Boltzmann 35
4.3.2.2 Generalized-Born 36
4.3.2.3 Solvent Accessible Surface Area 38

5 Molecular Dynamics Simulations **39**
5.1 Equations of Motion 39
5.1.1 Verlet 40
5.1.2 Leapfrog Verlet 40
5.1.3 Velocity Verlet 41
5.1.4 Langevin Dynamics 42
5.2 Ensembles 43
5.2.1 Berendsen Temperature Coupling 43
5.2.2 Pressure Coupling 45
5.3 Boundary Conditions 46
5.3.1 Free Boundary Conditions 46
5.3.2 Periodic Boundary Conditions 46
5.4 Force Fields 47
5.4.1 Bond Distance 47
5.4.2 Bond Angle 48
5.4.3 Urey-Bradley 48

5.4.4 Dihedral Angle 48
5.4.5 Improper Torsion 49
5.4.6 van der Waals 49
5.4.7 Electrostatics 50
5.5 Calculation of Non-Bonded Interactions 51
5.5.1 Spherical Cutoffs 51
5.5.1.1 Lennard-Jones Cutoff Schemes 51
5.5.1.2 Electrostatic Interactions 54
5.5.2 Ewald Summation 56
5.5.3 Non-bonded List 58
5.6 Advanced Sampling Techniques 59
5.6.1 Replica Exchange in Temperature Space 60
5.6.1.1 Free Energy Calculations 61
5.6.1.2 Scaling 61
5.6.2 Adaptive Biasing Force 62
5.6.2.1 Sampling 63
5.6.2.2 Error Calculation 64

6 Protein-Solvent Polarization in Myoglobin Hydration 65

6.1 Background 65
6.1.1 Myoglobin 65
6.1.1.1 Structure and function 66
6.1.2 Experiment 66
6.1.2.1 Hydration Number 68
6.1.2.2 Water Positions 68
6.1.2.3 Water-Protein Dynamics 69
6.1.3 Simulation 69
6.1.3.1 Protein Hydration 69
6.1.3.2 Water Positions 69
6.1.3.3 Water-Protein Dynamics 70

6.2 Methods . . . 70
6.3 Results and Discussion . . . 71
6.3.1 Dipole Moment . . . 71
6.3.2 Dipole Correlations . . . 71
6.3.3 Fluctuations, Deviations and Radius of Gyration . . . 74
6.4 Conclusions . . . 76

7 Molecular Dynamics Calculation of Microtubule Stability **78**
7.1 Introduction . . . 78
7.2 Orientational Restraint . . . 79
7.2.1 Quaternions . . . 79
7.2.1.1 Quaternions versus Rotation Matrices . . . 80
7.2.2 Minimizing RMSD . . . 81
7.2.3 Rotational Restoring Force . . . 81
7.2.4 Implementation . . . 83
7.3 Simulation Protocol . . . 83
7.4 Results and Discussion . . . 85
7.4.1 Protofilament Offset . . . 85
7.4.1.1 Convergence . . . 85
7.4.1.2 Flexural Rigidity . . . 87
7.4.1.3 Lattice Type . . . 90
7.4.2 Protofilament Separation . . . 90
7.4.3 Protofilament Bending . . . 92
7.5 Conclusions . . . 96

8 Conformational Analysis of Tubulin C-Terminal Tails **98**
8.1 Introduction . . . 98
8.2 Methods and Materials . . . 99
8.2.1 Parameterization and Model Preparation . . . 100
8.2.2 Cluster Analysis . . . 100
8.2.3 Principal Component Analysis . . . 101

8.2.4 Sequence Alignments 101
8.2.5 Motif Identification 101
8.3 Results 102
8.3.1 Amino Acid Composition of CTT Consensus Sequences 102
8.3.2 REMD and Completeness of Sampling 104
8.3.3 Clustering and Secondary Structure 105
8.3.4 Motif Identification 105
8.3.5 PCA 107
8.3.6 2D-Projections of the REMD conformation ensemble 107
8.3.7 Relative CTT Flexibility 108
8.4 Discussion 109
8.4.1 CTT Flexibility and Secondary Structure 111
8.4.2 Motifs and MAP interactions 113
8.5 Conclusions 113

9 Molecular Theory of Solvation **115**

9.1 Theoretical Background 116
9.1.1 Ornstein-Zernike Equation 116
9.1.1.1 Closure 117
9.1.1.2 Solvation Free Energy 118
9.1.2 1D-RISM 119
9.1.2.1 Solvation Free Energy 120
9.1.3 3D-RISM 120
9.1.3.1 Solvation Free Energy 121
9.1.3.2 Analytical Derivatives 121
9.2 Implementation 121
9.2.1 1D-RISM 122
9.2.2 Implementation and Optimization of 3D-RISM 122
9.2.2.1 Potential, Asymptotics and Force Calculations 123
9.2.3 3D-RISM Convergence 127

9.2.3.1 Solution Propagation 127
9.2.4 Multiple-Time Step Algorithms 127
9.2.4.1 Impulse Based MTS 128
9.2.4.2 Extrapolative MTS 128
9.2.4.3 MTS in Amber 129
9.2.4.4 Modified MTS for 3D-RISM in Amber 130
9.2.4.5 Coordinate Based Interpolative MTS 131
9.3 Free Energy of Solvent Polarization 132
9.3.1 Parameters for Site-Site Water Models 132
9.3.2 Free Energy of Solvent Polarization 132
9.3.2.1 Methods 134
9.3.2.2 Results and Discussion 134
9.4 Conclusions 135

10 3D molecular theory of solvation coupled with molecular dynamics in Amber 137
10.1 Introduction 137
10.2 Theory and Implementation 138
10.2.1 Implementation and Optimization of 3D-RISM 140
10.3 3D-RISM-KH and Molecular Dynamics Setup 140
10.4 Results 141
10.4.1 Net Force Drift Error 141
10.4.2 Energy Conservation 143
10.4.3 Optimization and MTS Speedup 144
10.5 Discussion 144
10.5.1 Factors Affecting Numerical Accuracy 144
10.5.2 Speedup 145
10.6 Conclusions 146

11 Summary and Future Work 147

A Amino Acids 151

B Protein Net charge 156

C CHARMM System of Units 159

List of Tables

3.1 Human tubulin isotype distribution and function 21

4.1 H_2O properties. 27
4.2 TIP and SPC water model parameters 34

8.1 ClustalW multiple sequence alignment of the CTT sequences 100
8.2 Sequence motif identification 103
8.3 PCA overlap of CTT peptides 105

9.1 Hydrogen radii for water and 3D-RISM 132
9.2 3D-RISM ΔG_{pol} of ALA_{10} 135
9.3 TI, 3D-RISM, PE and GB ΔG_{pol} of ALA_{10} 136

10.1 3D-RISM net force drift. 142
10.2 Constant energy decay rates. 143

A.1 Amino Acid Abbreviations and Properties 151

B.1 pK_a values for ionization of seven Amino Acids at 25° C [279]. 157

List of Figures

2.1 General structure of an amino acid 4
2.2 Primary structure of BPTI 5
2.3 Secondary structure of BPTI 5
2.4 Alpha helix 6
2.5 β sheet 7
2.6 Tertiary structure of BPTI. 8
2.7 Quaternary structure of hemoglobin. 8

3.1 Secondary structure of tubulin. 11
3.2 Tertiary and quaternary structure of tubulin. 12
3.3 MT structural highlights 13
3.4 GDP and GTP structure. 14
3.5 Microtubule lattice types. 15
3.6 Singlet, doublet and triplet microtubules. 15
3.7 Microtubule assembly and disassmebly. 17
3.8 Internal structure of cilia. 18
3.9 Mitotic apparatus. 19
3.10 Polar MT alignment and separation. 19

4.1 Phase diagram of water 26
4.2 Structure of H_2O. 27
4.3 Water-water hydrogen bonding. 28
4.4 Radial distribution functions of water. 29
4.5 SPC and TIP water model geometries. 33

5.1 Potential energy terms. 48
5.2 Comparison of Lennard-Jones potentials. 49
5.3 Lennard-Jones cutoff schemes. 53
5.4 Lennard-Jones cutoff schemes. 56
5.5 Ewald-charge distribution. 58

6.1 Secondary Structure of 1MBC. 66
6.2 Tertiary structure of 1MBC. 67
6.3 Evolutionary tree of the globulin family. 67
6.4 Electrostatic potential of myoglobin 72
6.5 Average dipole moment as a function of hydration 72
6.6 Dipole moments of surface water on myoglobin 73
6.7 Dipole-dipole and dipole-electric field correlations 73
6.8 RMSF of myoglobin 75

6.9 Average RMSF of myoglobin backbone 75
6.10 Time averaged RMSD of myoglobin heavy atoms. 76
6.11 Time averaged radius of gyration of myoglobin heavy atoms. 76

7.1 Reaction coordinates for MT interactions 84
7.2 Free energy profile of protofilament interactions along a longitudinal offset . . . 86
7.3 Protofilament-protofilament separation vs. longitudinal offset 86
7.4 Distortion of tubulin M- and N-loops 88
7.5 Taxol and epothilone binding sites 89
7.6 N- and M-loop interactions for A- and B-type lattices 90
7.7 Free energy profile of protofilament interactions along a lateral separation . . . 91
7.8 M- and N-loops at different lateral protofilament separations 91
7.9 Free energy profile of protofilament bending 92
7.10 S3-H3 interface 93
7.11 E-site conformational bending 94
7.12 N-site conformational bending 95

8.1 Illustration of reference and average structures used for PCA 102
8.2 ρ_{sc} distributions of CTTs about representative structures 106
8.3 Time average of each type of secondary structure for each CTT 106
8.4 Sequence motif identification 107
8.5 Root-mean-squared-normalized eigenvalues for each isotype 108
8.6 Projections of conformations onto the most important principal components . . 109
8.7 Peptides flexibility indicators 110

9.1 Amber and SANDER work flow charts. 122
9.2 3D-RISM flow chart. 123
9.3 Cutoff schemes for LJ and coulomb potentials and forces. 124
9.4 Experimental and 1D-RISM RDFs 133
9.5 SPC/E with alternate hydrogen Lennard-Jones parameters 133

10.1 3D-RISM MTS average temperature error. 143
10.2 Speedup resulting from interpolative MTS 144

Chapter 1

Introduction

Proteins are simple polymers that exhibit complex behaviour. Composed from a selection of 20 basic chemical building blocks, these molecules are machines that allow cells to function. While the contribution of these individual molecules to the complex functions in a cell can be understood through purely physical means, predicting their structure and function based solely on composition has been a difficult task.

The most powerful theoretical technique for the study and prediction of protein structure and function has been molecular modelling. Quantum mechanical, molecular mechanical, coarse-grained and mean-field methods have all been used for the study of protein folding, ligand binding and more.

In this thesis, we have used molecular dynamics to focus on three different topics in inter- and intra-protein interactions: water-protein interactions, microtubule properties and accelerating molecular dynamics through a mean-field solvation method. While these topics are diverse, they are all motivated by the greater goal of understanding protein interactions and their application to microtubule dynamics.

Background material can be found in Chapters 2 to 5. Chapter 2 provides basic background on the physical makeup and characterization of proteins. As such, it is an overview of the material this thesis will focus on. In Chapter 3, we focus on a particular protein, tubulin, and the structures it self-assembles to create, microtubules. Here the physiological properties and roles of microtubules are discussed and the current understanding of the molecular basis is presented. We then step back to look at water and how it influences protein structure and function in Chapter 4. The final introductory chapter details the background of our core methodological tool, molecular dynamics. Here the basic and advanced algorithms used in our various simulations are discussed.

The remaining five chapters each address different aspects of protein interactions and microtubule dynamics. This begins in Chapter 6, where the interactions of water, protein and their respective influences are investigated for the specific case of myoglobin hydration. Microtubule structure and stability is directly addressed in Chapter 7. A detailed picture of the molecular origin of microtubule properties, from its constituent protein tubulin, is developed through potentials of mean force calculated via molecular dynamics. Chapter 8 continues to investigate microtubule properties but focuses on sequence differences in the human body. The physical properties of the most variable part of tubulin, in both sequence and structure, is characterized using molecular dynamics. The enormous computational resources required for our investigations of tubulin motivate the implementation of a mean-field approach to molecular solvation in molecular dynamics. Chapter 9 details the theory of the 3D-reference interaction site model (3D-RISM) [1–4], implementation in the Amber molecular dynamics package [5]

and comparison against other models. Chapter 10 characterizes the computational performance of 3D-RISM coupled to Amber. In Chapter 11 we summarize the results of this thesis and contemplate the new questions it has led us to.

Chapter 2

Proteins

While we can not yet predict all of the physical properties of a given protein from its chemical composition, through decades of experimental and theoretical study, there is much we do know about the physical properties of proteins. This chapter serves to review the basic physical characteristics and terminology required to discuss protein structure and function. Section 2.1 describes the chemical composition of proteins. In Section 2.2, structural terminology is introduced and described. Section 2.3 provides a brief discussion of protein function. This basic knowledge, and much more, can be found in many sources such as [6] and [7]. It is from these texts that we have taken the material below except where otherwise noted.

2.1 Amino Acids

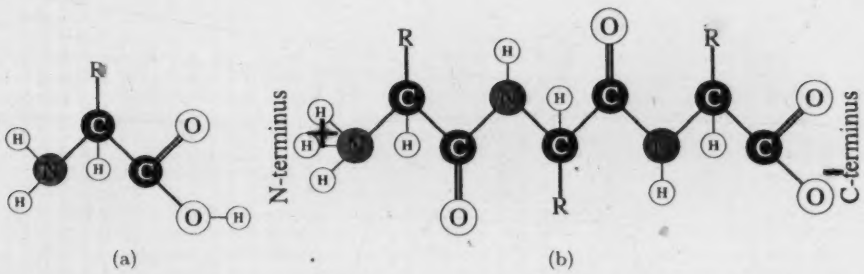
Individual proteins are single molecules, containing anywhere from 10's of atoms to 10's of thousands of atoms. We are fortunate that these large collections of atoms are constructed in a piece-wise manner out of chemical building blocks called amino acids (also called residues), of which there are only 20. However, these 20 amino acids each have a distinct atomic make up (see Appendix A), which is an alteration of a basic form (see 2.1(a)). It is with these building blocks, and sometimes embedded prosthetic groups, that proteins are made.

However, one should not make the mistake that the small number of standard amino acids imposes any serious limits on protein versatility. There are 20^n possible sequence combinations for a protein n amino acids long. For example, the protein myoglobin stores oxygen in muscles and is 153 residues long. It is one combination out of approximately 10^{199} possibilities. The vast majority of these other possibilities do not have stable conformations and are, consequently, not found in nature.

The basic structure of an amino acid is very simple and can be found in Figure 2.1(a). Every amino acid contains the three backbone atoms: N, C_α and C. The C_α also has bound to it a side-chain group that gives the amino acid its particular properties.

Amino acids chemically combine to form long polypeptide chains (Figure 2.1(b)). These chains have a direction due to the fact that the C of the backbone binds to the N of the next amino acid backbone, forming a peptide bond. As proteins are constructed starting at the N-terminus, sequences start at the N-terminus and end at the C-terminus, which are typically positively and negatively charged respectively at neutral pH.

Due to their side chains, amino acids fall into one of four different groups, acidic, basic, uncharged polar, nonpolar. A table of amino acids and their basic properties can be found in Appendix A. A discussion of the properties of acidic and basic amino acids can be found in



(a) (b)

Figure 2.1: (a) General structure of an amino acid. The backbone is made up, from left to right, of the nitrogen (blue), α carbon (blue) and carbon. The carbonyl oxygen is shown in yellow. "R" (red) represents the protein side chain. (b) Basic form of a polypeptide chain. This chain is made up of three residues. Residue 1 (side chain R_1) is the N-terminal residue. Residue 3 (side chain R_3) is the C-terminal residue. Residue 2 (side chain R_2) may be replaced by any number of residues.

Appendix B. These properties determine, in part, the location of the peptide with relation to the surface of the protein. Acidic, basic and polar groups are hydrophilic and tend to reside on the surface of the protein while nonpolar groups form the core.

Three amino acids that are often lumped in with the nonpolar residues are also known as "special" amino acids[6]. Cysteine contains a sufhydryl group (—SH) that can oxidize with a second cysteine bond to form a disulfide bond (—S—S—). This typically only occurs in extracellular proteins, to help maintain structure in diverse chemical environments. Glycine is the smallest amino acid as it has an side chain consisting of a single hydrogen. Finally, proline is unique in that the side chain is covalently bound to both the C_α and the N of the backbone (technically, it is an imino acid). This leads to a very rigid structure that often produces a fixed kink in the chain.

In summary, a protein is a long chain of specific amino acids in a specific order with its physical properties determined by the side-chains of the amino acids.

2.2 Structure

Despite typically having a globular, lumpy shape when "viewed" microscopically, proteins do have well ordered structures. In fact, there are generally four levels of structure in any given protein which are fundamental to the function of the protein. Changes in the structure of a protein on any of these levels may result in disease, such as sickle-cell anemia or Alzheimer's disease[6, 7]. Protein structure is governed by physical laws and is, therefore, predictable, though not easily so. The structure, in turn, defines how the protein physically interacts with other proteins and its physical environment.

2.2.1 Primary Structure

Primary structure is simply the sequence of amino acids in the polypeptide chain. This chain has a direction so residues are listed starting from the N-terminus. An example, bovine pancreatic trypsin inhibitor (BPTI)[8, 9], of such a structure is shown in Figure 2.2.

This sequence can be determined from DNA sequence that encodes it or from the protein directly. The protein folding problem, the task of determining the tertiary structure from the

RPDFCLEPPYTGPCKARIIRYFYNAKAGLC
QTFVYGGCRAKRNNFKSAEDCMRTCGGA
(a)

ARG PRO ASP PHE CYS LEU GLU PRO PRO TYR THR GLY PRO CYS LYS ALA ARG
ILE ILE ARG TYR PHE TYR ASN ALA LYS ALA GLY LEU CYS GLN THR PHE VAL
TYR GLY GLY CYS ARG ALA LYS ARG ASN ASN PHE LYS SER ALA GLU ASP CYS
MET ARG THR CYS GLY GLY ALA
(b)

Figure 2.2: Primary structure of (BPTI). BPTI consists of one polypeptide chain of 58 amino acids. (a) uses single character symbols for the residues while (b) uses three-letter codes. The N-terminus is located on the top left while the C-terminus is located on the bottom left for both.

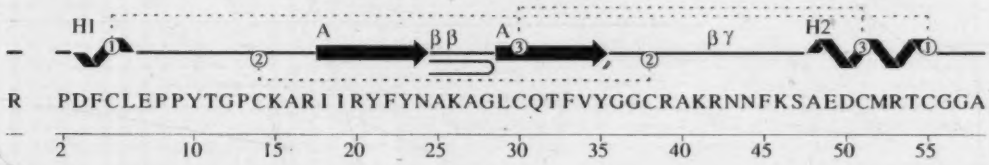


Figure 2.3: Secondary structure of BPTI. Helices are shown as coiled ribbons, β-sheets are arrows. Disulphide bridges between cysteine residues are labeled $1-3$. A hairpin turn where the β sheets loop back on each other is indicated by a red "U". β and γ turns are identified by β and γ symbols[10].

primary structure has been the focus of much attention.

2.2.2 Secondary Structure

The conformation and arrangement of nearby residues in a protein is known as the secondary structure. Individual residues are classified as having a particular type of secondary structure, the most common of which are α-helix or β-sheet. Other common types of structures are turns and different types of helices. Through x-ray crystallography over 100 types of secondary structure have been classified. The rest of the residues, typically 40% of the protein, is random coil and relatively flexible. A representation of secondary structure of BPTI can be found in Figure 2.3.

2.2.2.1 α Helix

α helices are rod-like structures formed by the backbone taking on a helical conformation (Figure 2.4(a)). This structure is stabilized by hydrogen bonds between carbonyl oxygens and the hydrogens bound to the backbone nitrogen, amide nitrogen, four residues along the sequence. This gives 3.6 residue per turn of the helix. Because the hydrogen bonds are aligned it gives the helix a net electrostatic polarization when the side chains are not considered.

Side-chains are excluded from the center of the helix. Since backbone polar groups are already involved in hydrogen bonds the hydrophobicity or hydrophilicity is completely determined by these side chains. Often hydrophobic side chains will extend to one side of the helix (toward the core of the protein) while hydrophilic ones will extend to the other. Such an arrangement is called amphipathic. This arrangement is common in globular and fibrous proteins

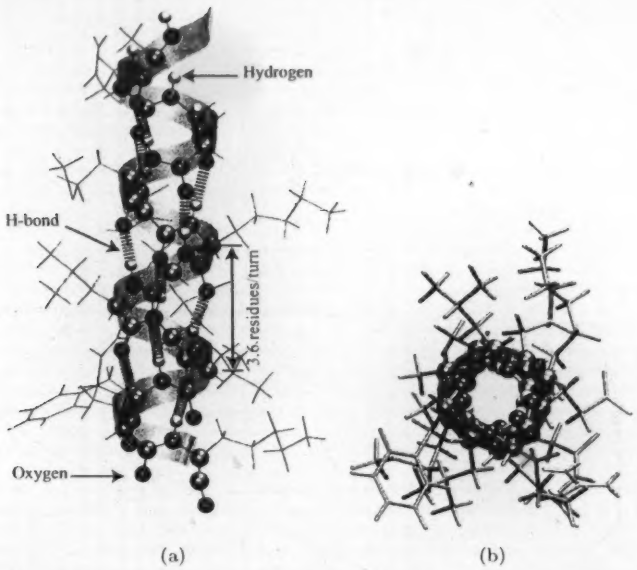


Figure 2.4: A subsection of a alpha helix from myoglobin. (a) The ribbon traces the backbone atoms of the protein which are shown as spheres. Oxygen and hydrogen are coloured red and yellow respectively while carbon and nitrogen are grey. Stabilizing hydrogen bonds between oxygen and hydrogen are coloured blue. Side chains are represented as lines, demonstrating how they are excluded from the center of the helix. (b) Hydrophobic residues are shown in orange while hydrophilic ones are shown in cyan. Images created with VMD [11].

that exist in a watery environment. An example of this can be found in Figure 2.4(b).

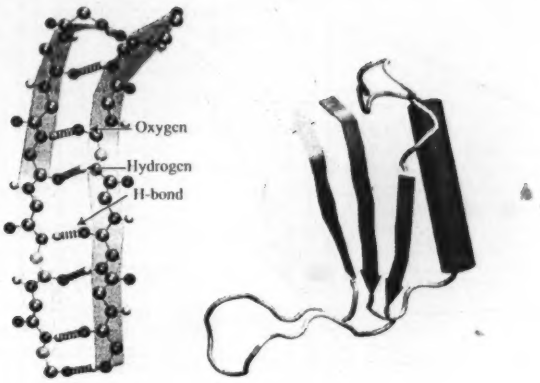
2.2.2.2 β Sheet

β sheets are the other most common structural element in proteins. Unlike α-helices, amino acids can form β-sheets with residues widely separated in the primary sequence. β-sheets are made up of two or more strands (continuous regions of primary sequence with their polypeptide backbone almost completely extended) laterally pack together such that amide hydrogens bond the carbonyl oxygens on adjacent strands (see Figure 2.5(a)). Each of these β strands is typically $5-8$ residues long and may be aligned parallel or anti-parallel (see Figure 2.5(b)).

The large number of hydrogen bonds makes β-sheets an extremely strong structure. For example, silk fibers consist almost entirely of stacked anti-parallel β-sheets. The polypeptide chain is aligned parallel to the fiber axis, which provides tensile strength. Flexibility is provided by the stacked β sheets slipping over one another.

2.2.2.3 Turns and Random Coil

Turns are commonly formed by three to four residues that, through hydrogen bonding, form a rigid U-shape that turns the peptide chain back into the protein. These turns allow the protein to become tightly packed. Common residues found in turns are glycine and proline. The small side-chain of glycine allows it to make sharp turns while proline already has a rigid built-in



(a) (b)

Figure 2.5: β sheet.(a) The ribbon traces the backbone and shows the direction of the peptide chain. Oxygen and hydrogen are coloured red and yellow respectively while carbon and nitrogen are grey. Stabilizing hydrogen bonds between oxygen and hydrogen are coloured blue. Side-chains are omitted for clarity. (b) β sheets are made of two or more contiguous sections of peptide chain that are not necessarily nearby in terms of primary structure. β sheet is orange, α helix is red, turns are yellow and random coil is grey. This piece of chain starts at the top of the middle β strand and ends at the top of the left β strand. Images created with VMD [11].

turn in it backbone.

Random coils are sections of the backbone with no rigid structure. These are typically found on the surface of the protein and tend to be very flexible. As a consequence, these also tend to be the areas that interact with other molecules. Typically, these are the sites of greatest conformational change after such interactions.

2.2.3 Tertiary Structure

The complete 3D structure of an amino acid sequence is known as its tertiary structure. Whereas secondary structure is stabilized through hydrogen bonds, tertiary structure is stabilized primarily by hydrophobic/hydrophilic interactions as well as salt and disulphide bridges. Nonpolar residues tend to collect in the center of the protein while hydrophilic residues line the surface as a general rule. In the case of membrane proteins the part of the surface embedded in the lipid bilayer also tends to be hydrophobic. As a result, the size and shape of a protein is determined both by the length of the amino acid and by how the secondary structure is arranged inside the protein. As an example, the tertiary structure of BPTI is shown in Figure 2.6.

2.2.4 Quaternary Structure

Proteins may be made up of one or more polypeptide chains. The number of these chains and how they fit together is called quaternary structure. Monomers, dimers and trimers are proteins composed of one, two and three chains, or subunits, respectively. These chains are held together by noncovalent bonds.

Hemoglobin, a classic example of quaternary structure, is shown in Figure 2.7. It is composed of four subunits, two α and two β. Each subunit contains a heme group that binds to O_2, CO_2 or CO since hemoglobin is used to transport oxygen in the blood.

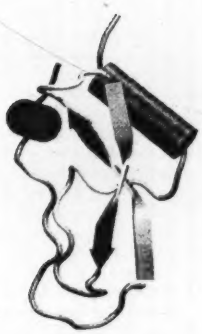


Figure 2.6: Tertiary structure of BPTI. β sheet is orange, α helix is red, turns are yellow and random coil is grey. The water accessible surface is shown as transparent. Image created with VMD [11].

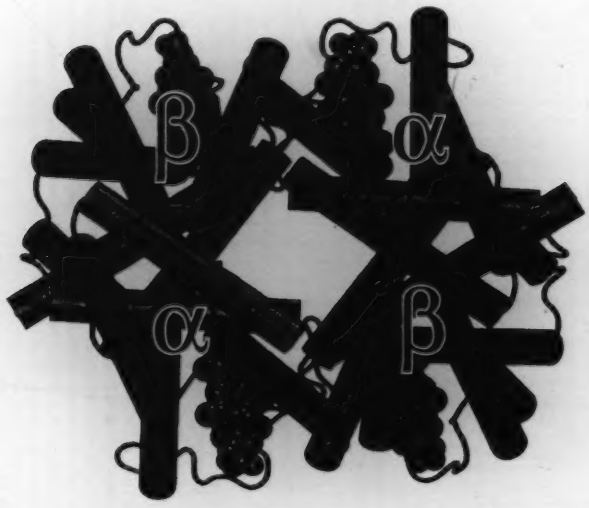


Figure 2.7: Quaternary structure of human hemoglobin (PDB ID: 1GZX)[8, 12]. α chains are shown in red and β chains are shown in blue. Image created with VMD [11].

2.3 Function and Role

Protein functions in the cell are many and varied, controlling the cell's ionic content, cellular locomotion, intra-cellular transport, replication of DNA, digestion of other molecules, etc... Common yeast, *Saccharomyces cerevisiae*, is thought to produce at least 6225 different proteins, each with a specific function. All of the elements of physical structure presented in the last section serve to determine an individual protein's function. Rather than focus on protein function in general, this section highlights the three topics of protein function found in this thesis.

2.3.1 Interaction with the Environment: Solvation

All proteins have evolved to function in a specific environment. Common to almost all these environments is their aqueous nature. Solvents are required for protein function and alter the structure of proteins. Likewise, proteins alter the structure of the hydrating water.

In Chapters 4 and 6 the quantity of water used to hydrate the protein myoglobin is examined. In particular, we investigate the amount of water required for myoglobin to attain functionality and how the properties of the water change as this limit is approached.

Chapter 4 also discusses methods for modelling solvent while Chapters 9 and 10 focus on one method in particular, the 3D-reference interaction site model (3D-RISM) of molecular solvation. Here we implement 3D-RISM in the molecular dynamics (MD) package Amber. Combining 3D-RISM and MD offers a new approach to modelling the complex interactions between solute and solvent.

2.3.2 Protein-Protein Interactions

Besides solvent, proteins most commonly interact with other proteins. In the case of tubulin, its primary function is to interact with other tubulins to form microtubules (MTs). Chapters 3 and 7 examines tubulin protofilament interactions essential for MT stability.

2.3.3 Conformational Flexibility

Proteins are not rigid objects and as they change conformation they change function. Two examples can be found in tubulin. Tubulin undergoes a conformation change that causes MTs to switch stable to unstable structures. In Chapters 3 and 7 the effects of nucleotide hydration on tubulin protofilament conformation and rigidity are examined.

A much smaller example, the tubulin C-terminal tails (CTT), are thought to change shape to bind to a variety of different proteins. In Chapter 8 and Section 3.3 we examine the conformational flexibility of the CTT and the isotype sequence dependence.

Chapter 3

Microtubule Structure and Function

Microtubules (MTs) are ubiquitous structures, found in all eukaryotic cells [6, 7]. These long, hollow cylinders carry out a wide variety of functions within the cell, including chromosome segregation, cellular locomotion and intracellular transport. This chapter provides detailed background on our current understanding of the structure and function of MTs, much of which is the subject of Chapters 7 and 8. Section 3.1, discusses the structure of the only element necessary to create MTs, tubulin. Section 3.2 covers how MTs are formed and the function of the resulting structure. Finally, Section 3.3 focuses on a highly variable part of tubulin's structure, the carboxy terminal tails, which have a profound effect on both the structure and function of MTs.

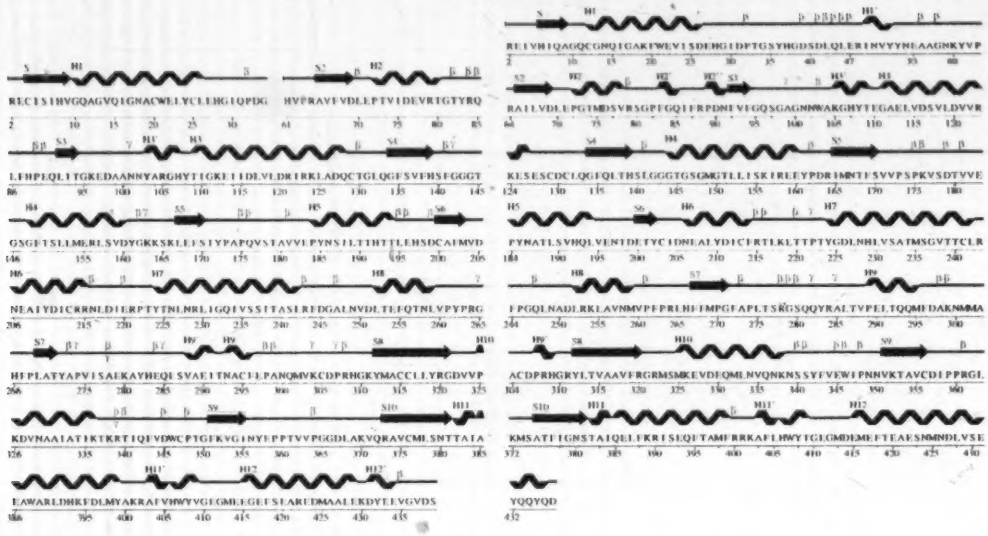
3.1 Tubulin

Tubulin is found in all eukaryotic cells [1]. Prokaryotic cells have a protein that is smaller, but still somewhat homologous to tubulin, called FtsZ. Tubulin is a dimeric protein consisting of an α and β monomer. The most common form of pig brain tubulin consists of 896 residues, 451 α residues and 445 β residues.

The crystal structure of cow brain tubulin was first published by Nogales *et at.* in 1998 (PDB ID 1TUB)[13]. The sequence for the most common form of pig brain tubulin was used for fitting the data since a sequence for cow brain tubulin was unavailable. The structure was subsequently refined by Löwe *et al.*, using the same data, and published in 2001 (PDB ID 1JFF)[14]. A number of different structures, under different conditions, have since been solved [15–18]. The final secondary structure (Figure 3.1) and 3D crystal structure (Figure 3.2) are missing several residues [2]. This is largely due to flexible parts of the chains not being resolved. These sections are both of the N-terminal residues, residues $35-60$ of the α chain and a large part of the C-terminal tails (CTT) ($440-451$ α, $438-455$ β). Although the CTT are the sites of greatest isotype variability (see Section 3.3.1.1), the poor resolution of the tails appears to be due to their flexibility (see Nogales *et at.* [13] and Section 3.3).

[1] Cells with nuclei, i.e. plant, animal and fungal cells. Tubulin, and MTs, are also found in red blood cells which have no genetic material and, thus, no nuclei.

[2] Additional gaps in the sequence also appear due to the convention used for tubulin residue numbering based on sequence homology [13]. Residues 45, 46 and $361-368$ of the β chain and 442, 443, 454, 455 of the α chain do not exist.



(a) (b)

Figure 3.1: Secondary structure of tubulin. The secondary structure of α tubulin, (a) and β tubulin, (b). Residues $35-60$ of the α subunit are missing. Helices are shown as coiled ribbons and β-sheets as arrows. Labeling has been adjusted to match that of [14]. β and γ turns are identified by β and γ symbols[10].

The primary function of tubulin in the cell is to self-assemble to form MTs. MTs are yet another layer of highly organized structure and are discussed in the next section. A third tubulin, γ tubulin, exists, has had its structure determined [17] and is thought to create nucleation sites for MT growth.

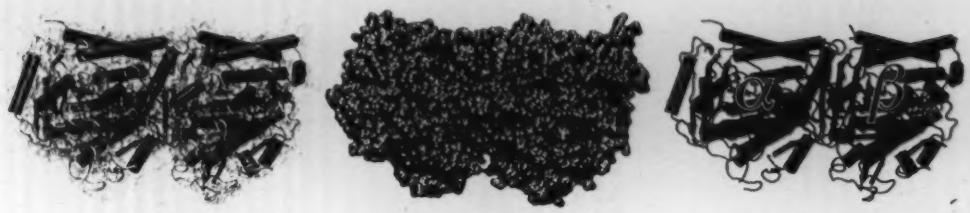
There are three areas of tubulin structure that are of particular importance for MT structure and function (Figure 3.3). These relate to lateral protofilament contacts, bound nucleotides and MT associated protein (MAP) binding sites.

3.1.1 MAP Binding

There are likely hundreds of proteins that interact with MTs directly but only a handful of binding sites are known. Through mutagenesis, proteolysis via substilisin and direct visualization, the CTT and the H11 and H12 helices have been identified as common binding targets for a variety of proteins [19–21]. This is not surprising given that these are highly exposed regions of tubulin when MTs are formed [22].

3.1.2 Lateral Contacts

Lateral contacts are primarily made between the M-loop (β-tubulin) and N-loop (α-tubulin) with the H3 helix [13, 14, 22]. These are sites of sequence variation in cold-stabilized MTs found in arctic fish [23, 24]. The M-loop has also been an issue for crystallographers. In crystals that have not stabilized the M-loop with taxol or epothilone, the M-loop has not been resolved [15, 16]. The N-loop has had no such issue, the equivalent of the taxol/epothilone



(a) (b) (c)

Figure 3.2: (a) Tertiary structure of crystallized tubulin. β sheet is orange, α helix is red, turns are yellow and random coil is grey. GTP, GDP and Mg^{2+} are shown in purple, cyan and green respectively. The water accessible surface is shown as transparent. (b) Water accessible surface. Hydrophobic residues are shown in orange while hydrophilic ones are shown in cyan. (c) Quaternary structure. α chains are shown in red and β chains are shown in blue. The orientation of all images is the same. Images created with VMD [11].

binding pocket on α-tubulin is filled with a 10 residue loop not present in β-tubulin. This is the primary evidence that taxol and epothilone stabilize MTs through stabilizing the M-loop and strengthening lateral interactions.

It has also been hypothesized that conformational changes in and adjacent to the H3 helix are responsible for destabilizing MTs. The H3′ helix and the S3-H3 loop are thought to resemble the highly conserved switch-II region of classical GTPases [14]. Furthermore, the H3′ helix is a π-helix in β-tubulin as compared to an α-helix in α-tubulin. This 'unwinding' of H3′ may be an indicator of real structural change or simply an artifact due to the poor resolution of the crystal structures.

3.1.3 Nucleotides

Tubulin also contains three non-covalently bound prosthetic groups. Each of the α and β subunits contains a guanosine nucleotide. The α subunit binds guanosine triphosphate (GTP, Figure 3.4(b)), which is not exchangeable, to the so-called N-site. The β subunit contains an exchangeable guanosine nucleotide, bound at the E-site, that may be either GTP or GDP (guanosine diphosphate, Figure 3.4(a)). At the N-site a Mg^{2+} is also noncovalently bound. Mg^{2+} has a high affinity to GTP and is always present at the N-site and at the E-site when GTP is present [25]. The lower affinity with GDP suggests that Mg^{2+} may be present when GDP is bound at the E-site but not necessarily and $Mg^{2+} \cdot$ GDP has been placed in some crystal structures [15].

3.2 Microtubules

MTs are constructed entirely out of tubulin. Along with microfilaments, intermediate filaments, MTs are one of the three major components of the cytoskeleton. Due to their tubular structure, they are the most rigid component and have a diameter twice that of intermediate filaments and three times that of microfilaments. MTs have a highly ordered and dynamical structure that is central to their function in the cell.

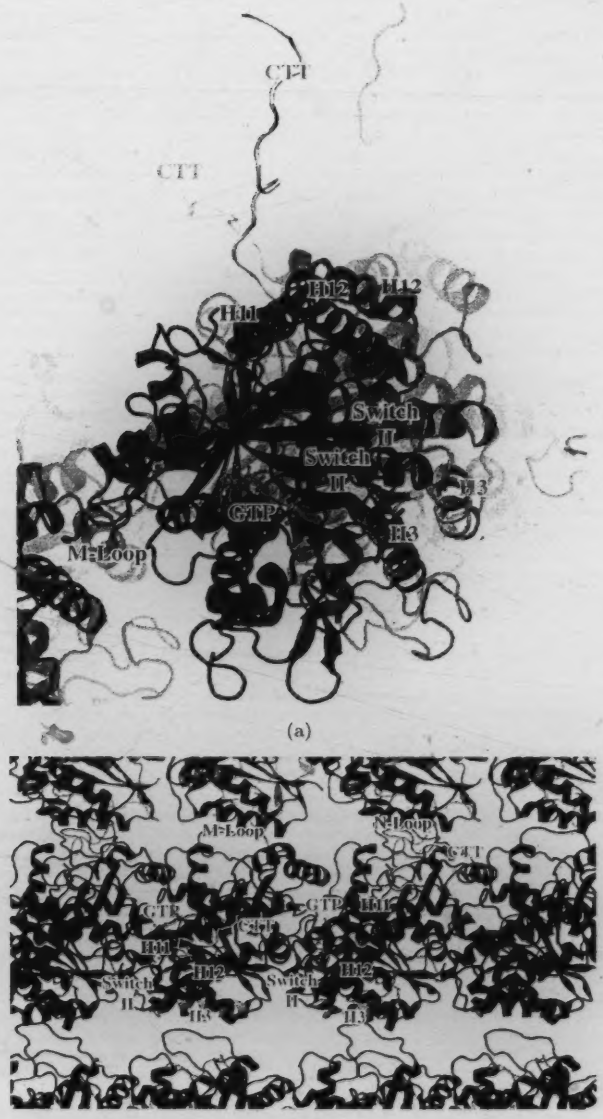


Figure 3.3: MT structural highlights (a) along the protofilament axis (E-site exposed) and (b) outside the MT. Images created with VMD [11].

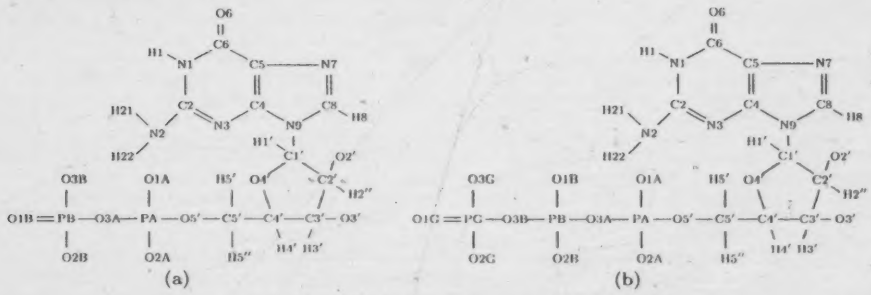


Figure 3.4: GDP (a) and GTP topologies (b) with IUPAC atoms names.

3.2.1 Structure

MTs are long, cylindrical structures with walls made out of an ordered array of tubulin dimers. The diameter of an MT is 25 nm with a 15 nm in diameter. Lengths vary from a fraction of micrometer to 100s of micrometers.

The wall of an MT lattice is of either type *A* or *B* [26–28] (see Figure 3.5). In both cases dimers are stacked longitudinally in protofilaments. There are relatively few examples of MTs *in vivo* that do not contains 13 protofilaments though *in vitro* this number may vary from 9 to 18 [26]. In *A* type lattices, each dimer is shifted 4.92 nm longitudinally from its neighbour. When there is an odd number of protofilaments, *A* lattices do not exhibit a seam. In the *B* lattice type each dimer is shift 0.92nm longitudinally relative to its neighbour, except at the seam, where the shift is 4.92nm. Longitudinally there is a new dimer every 8.12nm [14, 22, 26–28]. It was long believed that *A* lattices were preferred due to their symmetry. However, by labeling MTs with kinesin motor proteins, it was demonstrated that *B* lattices are, in fact dominant in number [27, 28].

MTs may be in one of three forms: singlet, doublet and triplet (see Figure 3.6). The singlet form is by far the most common and has already been discussed. Doublets take the form of a singlet of 13 protofilaments with an additional tubule of 10 protofilaments attached to the side. Triplets contain an additional tubule attached to the secondary tubule.

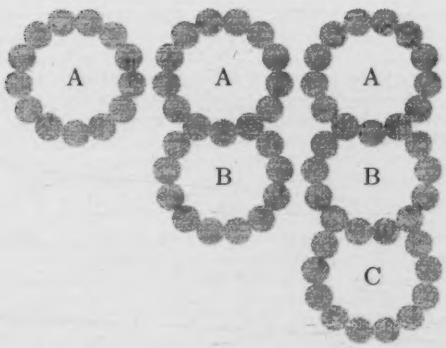
In vitro, tubulin structures can be quite diverse and are sensitive to changes in environment [29, 30]. The types of structures formed can be tubes, sheets, protofilaments, hoops, ribbons, double walled MTs etc. The three main variables for inducing different tubulin superstructures is the concentration of ions (mono- and divalent), anti- or pro-MT agents (colchicine, taxoids, vinca alkaloids etc) and the presence or absence of the carboxy terminal tails (Section 3.3.2.2). Some notable structures that can be created in this manner are tubulin sheets using zinc ions and taxol [13] and double walled tubes using manganese and partially subtilisin cleaved tubulin [31], both of which have been used to derive information about tubulin structure.

3.2.2 Dynamics

Tubulins self-assemble to form MTs. The physical characteristics and rates of polymerization depend on different environmental conditions, such as temperature, dimer concentration and GDP/GTP concentration. For example, assuming all other conditions are suitable, if the temperature falls to 4°C, yeast MTs will spontaneously disassemble [23]. However, if the

(a) (b)

Figure 3.5: (a) *A* type MT lattice has a 4.92nm longitudinal offset between lateral neighbours while the (b) *B* type MT lattice has a 0.92nm offset. α subunits are coloured red and β subunits are colour blue. Images created with VMD [11].



(a) (b) (c)

Figure 3.6: (a) Singlet, (b) doublet and (c) triplet MTs. Adapted from [6].

temperature is raised above this temperature the MTs will reform [3].

The process of self-assembly begins when the temperature is in the right range and the concentration of dimers and GTP are above the critical concentration, which is environment dependent. If nucleation sites, such as already formed MTs or γ tubulin, are present then the process is accelerated. Since tubulin has a polarization due to its quaternary structure the MT also has a polarization. The (-) end has exposed α monomers while the (+) end has exposed β monomers with exposed E-sites. Both MT assembly and disassembly occur preferentially at the (+) end.

It is not definitively known how γ tubulin acts to nucleate MTs though two models exist[32–34]. The first suggests that γ tubulin forms a ring or partial ring through lateral contacts. $\alpha\beta$ dimers then longitudinally bind to this ring. A second model suggest that the γ tubulin forms rings or curls through longitudinal contacts and that $\alpha\beta$ dimers then bind laterally in sheets. However, it is known that γ tubulin caps the (-) end of the MT and prevents both assembly and disassembly.

Electron micrographs of the γ-tubulin ring complex (γ-TuRC) are strong evidence that the template model at least exists [35]. There is no similar structural evidence for the protofilament model. However, the fact that not all eukaryotes, such as budding yeast [34], have γ-TuRCs, but only γ-tubulin small complexes (γ-TuSC), one of the subunits of γ-TuRC, strongly suggests that this model is still a possibility. In favour of the template model is the crystal structure of γ-tubulin, which indicates that it can form lateral bonds itself and γ-TuSCs may act as templates alone [17].

MTs grow through individual protofilaments adding tubulin dimers [36]. Only tubulin dimers with GTP in the exchangeable site are able to polymerize. However, once a dimer has spent some time in the MT it hydrolyzes the GTP to GDP+PO_4. Thus, a GTP cap is formed on the end of the MT as shown in Figure 3.7(a). However, if conditions for the growth of the MT are poor, the rate of tubulin addition is slower than the rate of GTP hydrolysis and the cap disappears. At this point rapid shrinkage occurs, known as catastrophe. When catastrophe occurs, the end of the MT frays and the protofilaments peel off, maintaining their longitudinal contacts but losing the weaker lateral ones. When conditions become favourable again the GTP cap will reform and the MT with grow again (rescue). This process, outlined in Figure 3.7, is called dynamic instability.

Another related phenomenon is treadmilling. In this case, conditions favour assembly at the (+) end but favour collapse at the (-) end. Thus, the MT grows at the (+) end and shrinks at the (-) end and dimers joining the MT at the (+) end will, eventually, leave at the (-) end.

There are several molecules that can stabilize MTs and prevent their disassembly [37]. Microtubule-associated proteins (MAPs), such as MAP2 and Tau, may bind to two or more dimers, acting as a support or neutralizing the negative net charge. Other MAPs like CLIP-17 and EB1 may copolymerize with new tubulin subunits and induce a particular conformation. An important small molecule that also stabilizes MTs is taxol. Taxol was instrumental in the crystalization of tubulin to determine its structure [13]. It is also a widely used therapeutic agent against various cancers [38].

Other molecules are known to disrupt MTs. Once again, they may be proteins (e.g. kinesin 13 [39]) or small molecules. Some of these may also have therapeutic applications as well (e.g. colchicine in gout [15, 40] and vinblastine in Hodgkin's disease [16, 38, 41]).

[3] Temperature dependence on MT polymerization is known to be isotype dependent [23, 24]. Some arctic fish MTs can polymerize at temperatures as low as -1.8°C. There is even considerable variation amongst human isotypes.

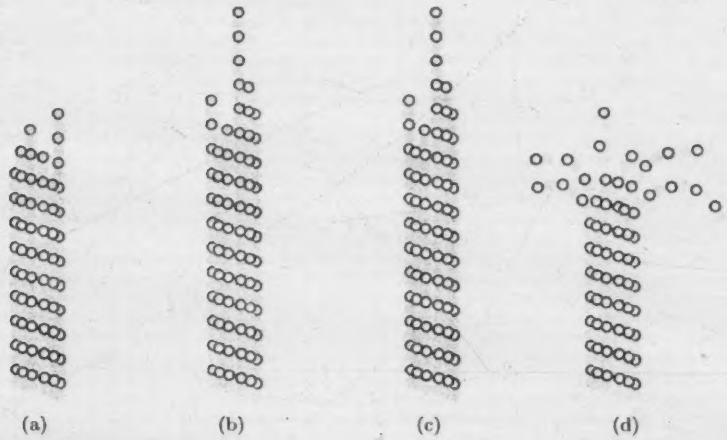


Figure 3.7: Microtubule assembly and disassmebly. α subunits are coloured light blue and β subunits dark blue. GTP and GDP are represented by light blue and yellow centres respectively. (a) A GTP cap is formed on the MT allowing (b) further growth. (c) GTP hydrolysis occurs faster than growth causing the GTP cap to disappear and (d) catastrophe to occur. At some future time the GTP cap reforms and the cycle begins at (a) again. Adapted from [36]

3.2.3 Function

MTs perform a variety of functions within the cell. The three predominant tasks performed by MTs all involve motion: transportation of molecules within the cell, locomotion of the cell in its environment and cell division.

3.2.3.1 Intracellular Transportation

MTs are the railway system of the cell. They provide the tracks on which molecules like kinesin and dynein haul their molecular cargo [7]. Nowhere is this as important as in nerve axons. The cell body of the nerve, where the ribosomes are present, can be several meters from the synapse where proteins, mitochondria and membranes are required. While this process requires days or weeks with motor proteins, it is much faster than simple diffusion, which would require on the order of a decade.

This transport is aided by the fact that MTs grow out from the MT-organizing center (MTOC), located in the centrosome. That is, their (-) ends point toward the nucleus of the cell while the (+) ends radiate outward. Kinesin and dynein are directional motor proteins. Almost all isotypes of kinesin move toward the (+) end while dyneins move toward the (-) end. Different types of each of the respective proteins carry a specific load. Cytosolic kinesin carries cytosolic vesicles (closed membrane shells). Spindle kinesin carries spindle and astral MTs, centrosomes and kinetochores. Cytosolic dyneins carry cytosolic vesicles and kinetochores.

Axonal transport can be very rapid; motor proteins can travel over 300 of their own body lengths per second. This is about 3 μm/s or 250 mm/day. Thus, for a typical sciatic nerve of a human, approximately 1m in length, the fastest axonal transport can take nearly 40 days.

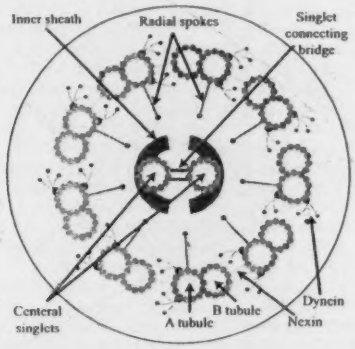


Figure 3.8: Cross section of the internal structure of cilia. The inner sheath with a pair of MT singlets is connected to the outer ring of doublets by a set of radial spokes. Outer doublets are connected to each other by the protein nexin. Inner and outer dyneins are permanently attached to the A tubule and walk along the adjacent B tubule. Adapted from [6].

3.2.3.2 Cellular Transportation

Cilia and flagella [4] are the common method for swimming for all cells [7]. Their function, however, is not limited to swimming. Some cells have one or two long flagella (up to 2 mm long in the case of some insect sperm) or may have thousands of short cilia used for feeding.

The internal structure of flagella is shown in Figure 3.8. Typical structure of the axoneme has two central singlet MTs and nine outer doublets. In addition to α and β tubulin there are over other 250 polypeptides present. Among these is dynein, which permanently binds to the A tubule of a doublet with its motor protein heads reaching out to the B tubule of the adjacent doublet. Other proteins, such as nexin, stabilize the axoneme, connecting the doublets to each other and the central singlets, and providing a central sheath around the singlets.

Cilia movement is caused by the movement of the outer doublets relative to each other. In particular, the dynein molecules 'walk' along the adjacent MT, pushing its (+) end away from the cell and bending the axoneme. Permanent crosslinks pull the MTs back to their original place.

3.2.3.3 Cell Division

The primary role of MTs in cell division is the partitioning of newly replicated chromosomes into, what will be, two separate cells [7]. Figure 3.9 shows a schematic snapshot of the cell and the mitotic apparatus during mitosis. It should be noted that the mitotic apparatus is anything but static and dynamic instability, discussed in Section 3.2.2, is central to the success of this stage in the cell cycle.

Durfng prophase, polar MTs are responsible for first aligning and then separating the centrosomes (Figure 3.10). The mechanism involved in the moving the MTs is similar to that of movement of cilia and flagella. Motor proteins attached to one MT walk along an adjacent MT. During alignment (-) end-directed motor proteins change angle of the MTs, taking up the slack. Once aligned, (+) end-directed kinesins push the MTs, and with them the centrosomes, away from each other.

[4] Cilia and flagella are the same structures but were named before their structure could be studied.

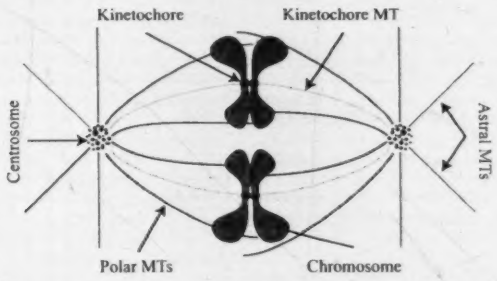


Figure 3.9: The mitotic apparatus is composed of centrosomes (red dots), astral MTs (orange), polar MTs (green) and kinetochore MTs (yellow). Adapted from [6].

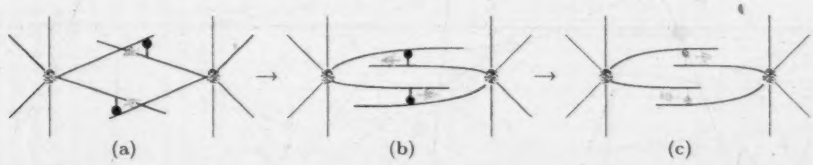


Figure 3.10: (a) Misaligned polar MTs are aligned by (-) end-directed motor proteins (blue). (b) These same motor proteins then pull the two centrosomes closer together. Once completely aligned (+) end-directed motor proteins push the centrosomes apart again. Adapted from [6].

MTs begin interacting with the chromosomes during prometaphase, when the nuclear envelope breaks down. Two centrosomes (organizing centers for the MTs) have already formed and the (+) ends of the MTs can now interact with the chromosome kinetochores.

At metaphase there are two centrosomes at either end of the cell and the chromosomes have duplicated but sister chromosomes are still attached to each other. The mitotic apparatus has formed and consists of astral, polar and kinetochore MTs.

Kinetochore MTs control the alignment and separation of the chromosomes. First, chromosomes are gathered via their kinetochores attaching to the kinetochore MTs. Attaching to the kinetochores is one example of how dynamic instability is used in the cell. There are several theories of how this is done but most involve dynamic instability or treadmilling. One theory is that the MTs grow and shrink, fishing for the kinetochores. An MT may strike a kinetochore but it is more likely that the kinetochore will attach to motor proteins on the surface of the MT and be pulled to the (+) end. After attachment the chromosomes are moved to the center of the cell.

The sister chromosomes separate during anaphase and the kinetochore MTs pull the chromosomes back into their respective centrosomes.

Finally, the cell divides. This occurs on the plane that the chromosomes aligned on in the center of the cell. The action of dividing the cell is caused by the contraction of an actin and myosin ring but the location of this cleavage appears to be determined by the astral MTs alone.

Since dynamic instability plays a central role in cell division and the mitotic apparatus, drugs that stabilize MTs can prevent cell division [38, 41]. This is how drugs like taxol can be used on cancers, such as ovarian cancer in which the cells rapidly divide, without overly affecting the other roles of MTs within the cell.

3.3 Carboxy-Terminal Tails

A mounting body of evidence suggests that tubulin's carboxy-terminal tails (CTTs) have a wide and varied biological significance. CTTs typically account for 20 of the roughly 450 residues that make up a tubulin monomer; however, they are the sites of both the greatest charge density and isotypic variation in the tubulin sequence. With only 50-60% sequence identity, compared to 80-95% in the total sequence [42, 43], we can, at least, speculate that this region has been under specific evolutionary pressure to have different physical properties for different biological functions. Much of the conserved sequence identity between isotypes is in the form of glutamic acids that account for typically half of the residue number, 1/3 of the net charge of tubulin and are the source of the CTTs less commonly used name, 'E-hooks'.

In this Section, we begin with a brief description of the physical structure of CTTs and variations due to sequence differences and post-translational modifications. With this background, we then discuss the many biological roles of CTTs and how physical variations are important in them.

3.3.1 Physical Properties of CTTs

CTTs are highly flexible and have no intrinsic structure in isolation [44–46] or in their native state, attached to tubulin [13]. This high flexibility lends them the ability to bind or interact with a large number of substrates. Due to their location on the outside of the MT, they are, inevitably, encountered by all proteins seeking to bind to MTs (Figure 3.3).

Though CTTs are often thought of as floppy, amorphous, highly charged protrusions, there can be considerable variability in physical properties within a single cell. The length, charge and binding motifs of an individual CTT is determined first by its primary sequence (isotype) and then modifications made after expression (post-translational modification).

3.3.1.1 Isotypes

The number of isotypes expressed within a cell or an organism varies greatly from species to species. Here, we focus on humans, having eight α and eight β isotypes though some subtypes also exist. Due to the high sequence variation of the CTT, isotypes are generally identified by their CTT sequence. Known human tubulin isotypes and their basic distribution and function are given in Table 3.1.

3.3.1.2 Post-translational Modifications

Post-translational modifications (PTMs)are changes made to side-chains or entire residues after the individual protein has been expressed. In the case of tubulin, almost all PTMs are applied to the CTT. Residues may be added or deleted. Side-chains may modified and are often extended. Below is a summary of two excellent review papers [48, 49].

Poly-Glycylation Poly-glycylation is the covalent addition of glycine residues to glutamic acid side-chains. In particular, this is a peptide bond between the γ carboxyl group of the glutamic acid and the α amino group of the glycine. This can occur on adjacent glutamic acids and 30 or more glycines may be attached in a single chain, making for a very bulky CTT. It is typically found only in animal cells and predominately in axonemes. The modification is thought to confer stability but the evidence is not clear.

Isotype	Function	Distribution
	β-Tubulin	
I	Found in all tissue types but varies in amount May confer stability	Widespread
II	58% of bovine brain tubulin Expressed more during development Function is not well understood	Brain, muscle
III	Highly conserved (two residue difference between human and chicken squence) In the brain, only found in neurons Accounts for 25% of bovine brain tubulin Found in some cancers Very dynamic	Neurons, Sertoli, testis, colon
IV	IVa is found only in the brain IVb is found in cilia and flagella IVb only isotype with the EGEFEEE sequence	Brain, ciliated tissues, sperm
V	Highly conserved Unknown function	Unknown
VI	Least conserved	Platelets, bone marrow,spleen
VII	Unknown function	Brain
VIII	Unknown function	Unknown
	α-Tubulin	
I	Unknown function	Mostly brain
II	Unknown function	Testis
III	Unknown function	Brain, muscle
IV	Unknown function	Blood
V	Unknown function	Heart, skeletal muscle, testis
VI	Unknown function	Testis
VII	Unknown function	Ovary
VIII	Unknown function Lowest sequence identity	Heart, testis, skeletal muscle, brain and pancreas

Table 3.1: Humān tubulin isotype distribution and function [42, 43, 47].

Poly-Glutamylation This is a similar PTM modification to poly-glycylation. Glutamic acids are added in chains to the γ carboxyl group of the main chain glutamic acid side-chains. In this case, the number of residues in a single chain is typically one to six. While this is far less bulky than poly-glycine PTMs, it can make the already negative CTT significantly more negative. Chains of three to four are associated with greater MAP binding and, possibly, greater MT stability. It is also observed to increase processivity of kinesins. Poly-glutamylation is particularly prevalent during mitosis.

(De-)Tyrosination This PTM occurs only on the C-terminal tyrosine of α-tubulin, which is present in almost all isotypes. In this case, the entire residue is removed and is, typically, replaced later. In fact, this may happen multiple times in the lifetime of a tubulin. The function is not known but their is a strong correlation with decreased de-tryosination and most breast cancers. Cells that lack the de-tryosination enzyme are known to cause tumors when injected into mice.

De-Glutamylation If the terminal tyrosine has been removed from α-tubulin, the terminal glutamic acid may also be removed. Once this occurs, neither the glutamic acid nor the tyrosine is replaced. 35% of mammalian brain tubulin is de-glutamylated and this is thought to confer MT stability.

Phosphorylation Phosphorylation is a relatively rare PTM. It occurs only in mammals, may aid the interaction with MAP2 and does not affect assembly.

3.3.2 Biological Role

CTT are implicated in a wide variety of biological processes. Most are are associated with MT functionality but some functionality beyond this has been proposed. For example, the CTT have been shown to confer chaperone-like activity to tubulin, preventing the aggregation of insulin and alcohol dehydrogenase [50]. Also, CTT peptides have been shown to form amyloid fibrils in vitro [51]. These have been directly implicated as the source of amyloid plaques found in the brains of British type familial cerebral amyloid angiopathy victims.

This section discusses the two most common CTT associated functions: tubulin polymerization and MAP binding. This is followed with a discussion of the role of the CTT and the binding of two common anti-mitotic agents. Most of the findings presented here were made by removing one or both of the CTT from tubulin. The only known method for doing this is with the enzyme subtilisin, which is where we begin.

3.3.2.1 Subtilisin

Subtilisin is a general class of serine proteases commonly expressed by bacteria and is the only enzyme known to selectively cleave the CTT from tubulin [52, 53]. It uses the same mechanism as chymotrypsin and serine carboxypeptidase II but has no sequence or structural homology with either of these other classes. In fact, other than the functional core of the protein, there is not much sequence identity within subtilisin class [52].

CTT proteolysis is a time dependent process [54, 55]. First, β is cleaved at Gln-433 and Gly-434. Some time later, α tubulin is cleaved at Asp-438 and Ser-439. However, there is some overlap in the two stages and contamination between the two species can easily occur without proper procedures [54]. The standard nomenclature is 'tubulin$_S$' for tubulin without CTT (alternatively '$\alpha_S\beta_S$'). If only the β CTT has be removed, then the notation '$\alpha\beta_S$' is used.

3.3.2.2 Tubulin Polymerization

One of the immediate observations made when subtilisin was first added to tubulin in solution was its effect on polymerization [56]. The critical concentration for polymerization is lowered by one to two orders of magnitude and appears to additively increase as first the reactions proceeds from $\alpha\beta_S$ to $\alpha_S\beta_S$. Also, very small amounts of $\alpha\beta_S$ added to native tubulin can profoundly affect the polymerization [54]. However, few MTs are formed from this polymerization. Rather, spirals, sheets, tubes, rings, bundles and short protofilaments are observed. The increase in polymerization is thought to result from the reduction of the total charge of tubulin as similar structures and critical concentrations are observed with the increased concentration of salt [29, 30].

3.3.2.3 MAP Binding

Any protein that binds to MTs is classified as a MT associated protein (MAP). Here we focus on two with well studied MAPs, kinesin and spastin, and their proposed interactions with tubulin's CTTs.

Kinesin Binding While kinesins come in many flavours, there are two broad classes of this motor protein, monomeric and dimeric (conventional) kinesi, that interact with CTTs in a fundamentally different ways. The 'heads' of kinesin bind to the β monomer of tubulin and walk towards the (+) end, hydrolysing adenosine triphosphate (ATP) as a fuel. Both types are known to require CTT to process along the MT.

Dimeric kinesin was thought to interact with the CTTs through a lysine rich region in the neck linker, using electrostatic interactions to strengthen affinity. This was supported by experiments that showed greatly reduced processivity when CTTs were absent and kinesin's with increased lysines content in the neck linker displaying increased processivity. However, the current consensus is now quite different [57]. Cryo-electron microscopy (cryo-EM) and various other experiments have shown that in the usually weakly bound adenosine diphosphate (ADP) state of kinesin becomes strongly bound when CTT are absent [58, 59]. That is, kinesin gets stuck in the ADP state and the role of the CTTs is to reduce the binding affinity, promoting the next step. Further, cryo-EM has shown that the CTT likely bind to the switch-II region of kinesin and not the neck linker. When experimenting with NcKin, a fast fungal kinesin with no lysines in the neck linker, Marx *et al.* found the same property [57]. When CTT were absent, kinesin stalled in the ADP state. This is also consistent with observation of high salt concentrations that would screen CTT-kinesin interactions.

Monomeric kinesin appears to use the β CTT as an anchor. KIF1A has a so-called k-loop (high lysine content) that appears to bind the β CTT [60, 61]. This k-loop is not present in dimeric kinesin. In fact, these cryo-EM electron densities have been the only experimentally observed density of CTTs. Unfortunately, these where not at atomic resolution, so details of the interaction are not known.

Spastin Binding Spastin is a large, star shaped hexamer implicated in MT severing. Defects in the subunits lead to hereditary spastic paraplegia [20]. The recent crystal structure of this protein shows a central pore hypothesized to thread CTTs through, pulling apart MTs. Spastin has a preference for severing poly-glutamylated MTs that are otherwise quite stable. This suggests a roll in specifically de-constructing stable MTs when they are no longer needed.

3.3.2.4 Drug Interaction

Colchicine Colchicine is an anti-MT agent that binds irreversibly to tubulin after inducing the formation of its own binding pocket [15]. In native tubulin, the binding site is labile for 4-5 hrs and there is a pH dependence for binding, with an optimum of 6.8 [62]. When both CTT are removed with subtilisin tubulin remains labile for $>$ 12 hrs and the pH dependence on binding vanishes. A similar effect can be produced with the addition of cations, suggesting that the α CTT interacts with the colchicine binding site.

Vinblastine Vinblastine has a more complex relationship with tubulin [55]. At concentrations $< 1\,\mu$M vinblastine inhibits MT dynamics. When increased above $1\,\mu$M but $< 10\,\mu$M it inhibits MTs. As concentrations are further increased above $10\,\mu$M, vinblastine begins to promote polymerization. Rather than forming MTs though, tubulin forms spirals and other polymers. The effects of vinblastine can be blocked with the addition of oligoanions, such as GTP. However, if the β CTT is removed, polymerization by vinblastine is increase and the concentration of oligoanions must be further increased to block this behaviour. Since further removing the α CTT has no effect, it is thought that the β CTT's negative charge blocks the binding of vinblastine.

Chapter 4

Protein Hydration

An accurate physical description of a protein molecule is not enough to understand its function or structure. Proteins have no biological functionality without water and require a proper solvent environment to properly fold. To study proteins in detail it is essential to have an accurate description of water and its interactions with protein. This chapter begins with an overview of the physical properties of water in Section 4.1, our current understanding of protein-water interactions in Section 4.2 and some of the most popular water models used in the study of proteins in Section 4.3.

4.1 Water

Water is the essential solvent for life on earth. Because of this, water has been the most studied molecule and has developed, in general, a certain amount of lore if not mystique. Certainly, much has been made of the so-called anomalous properties of water. While none of these individual properties is unique to water, there combination is exploited by life on Earth, particularly at the microscopic level.

To accurately model the properties of proteins and their interactions it is necessary to also include the effects of water. The focus of this Section is to identify the biologically important properties of water and their molecular basis. This should serve as a guide for creating and assessing models of this ubiquitous solvent.

Material for this Section is taken from three review papers by Finney [63, 64] and Guillot [65]. Additional sources are otherwise noted.

4.1.1 Macroscopic Properties

There are a number of properties of water that are necessary for life. In short, water is

- a liquid at standard pressure and temperature,
- an excellent solvent for a wide variety of substances and,
- able to ionize many substances, including, but not limited to, salts.

These properties relate to water's role as an almost universal solvent: dissolving salts, sugars, acids, bases, gases, proteins etc. Many of these substances, like salts and acids, are further ionized by water. This is possible because, unlike most simple liquids, water is a liquid at ambient temperature (Fig. 4.1).

Electrostatic properties of water include

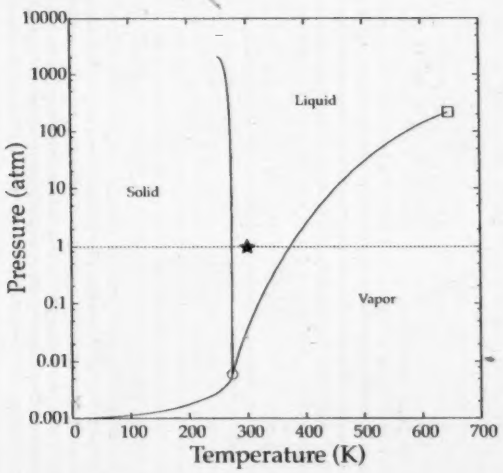


Figure 4.1: Partial phase diagram of water. ◯ indicates the triple point, □ the critical point and ★ is room temperature at ambient pressure. Coexistence curves fit to experimental data from the CRC Handbook [66].

- having a high dielectric constant,
- being a good insulator but also
- being a good proton conductor.

While these are bulk properties they are most important at the microscopic level where they mediate interactions between molecules.

At the organism or environmental level, we can also add that water has

- a higher density as a liquid than a solid and
- a high specific heat capacity.

Certainly, colder ecosystems largely depend on ice floating on water. Also, temperature regulation for individual organisms - and the planet as a whole - relies on the high specific heat capacity of water.

While this list is not exhaustive and none of these properties are unique to water, all are important and can be accounted for at the molecular level.

4.1.2 Molecular Properties

To explain the macroscopic properties of water we need to examine the microscopic properties. This starts with the geometric structure and electrostatic properties of the molecule. In turn, this gives water it's dipole moment, polarizability and hydrogen bonding properties. Particularly, from hydrogen bonding, we explain the microscopic structure of bulk water and many of the macroscopic properties of water.

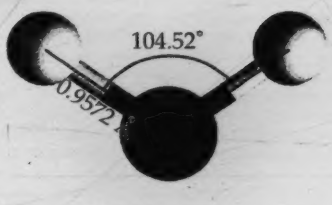


Figure 4.2: Structure of H_2O with gas phase bond lengths and angles. Oxygen is shown in red and hydrogen in white. Image created with VMD [11].

Property	Gas	Liquid
Bond length (Å)	0.9572	0.925-1.0[a]
Bond angle (°)	104.52	
Dipole (D)	1.85	2.6-3.0
Polarizability ($Å^3$)	1.45	1.3

Table 4.1: H_2O properties in gas and liquid phase.[a]For D_2O [68].

4.1.2.1 Intramolecular Structure

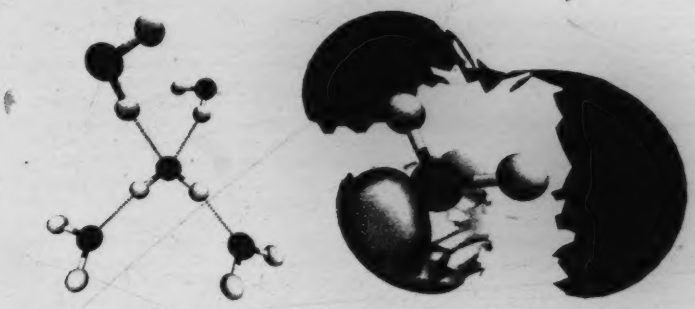
Fig. 4.2 gives geometric structure of water and Table 4.1 parameters of the structure in liquid and gas phase. The H-O bond length is 0.9572 Å and the H-O-H angle is 104.52° as measured through gas phase spectroscopy [67]. Note that this angle is very close to the tetrahedral angle (109.47°) and the internal pentagon angle (108°). Bond length and angle values have been difficult to measure in the liquid phase. Recent work on liquid D_2O found this bond length stretched from 0.925 to 1.0 Å depending on distance from neighbouring waters [68]. It should be noted that the asymptotic value for long range separations is lower than the gas phase bond length measured for H_2O. The electrostatic potential is polarized over this structure with the positive charge localized towards the hydrogens and negative around the oxygen.

4.1.2.2 Dipole Moment

As the net charge of the molecule is zero, the dominate term in the multipole expansion of the potential is the dipole. In the gas phase this has been experimentally measured to be 1.85 D. The answer is not clear for liquid water though. Both experiment and *ab initio* calculations put the average value around 2.6-3.0 D but suggest that individual dipoles may very in the range of 2-4 D. Of course, higher order terms in the multiple expansion also contribute, and short range interactions are not well described by a simple dipole.

4.1.2.3 Polarization

As suggested by the difference between gas phase and liquid phase dipole moments, water is significantly polarizable. *Ab initio* calculations suggest that the polarizability is about 10% less in liquid water than gas phase [65]. This susceptibility to local electrostatic fields contributes, in part, to two of water's characteristic properties: its dielectric constant and ability to ionize solutes.



(a) (b)

Figure 4.3: Water-water hydrogen bonding. Colouring as in Fig. 4.2. (a) Water from a simulation of TIP3P water [69] with hydrogen bonds (dashed lines) to four neighbours. (b) Isosurface of hydrogens (white) and oxygens (red) around a single SPC/E water [70] calculated by 3D-RISM [3, 44]. Images created with VMD [11].

4.1.2.4 Hydrogen Bonding

A single water can make a total of four hydrogen bonds (HBs); hydrogens act as donors and the two oxygen lone pair sites act as acceptors. Fig 4.3 shows a snapshot of a single water with HBs to its four nearest neighbours and the average positions of water hydrogens and oxygens around a water molecule. This 2-and-2 HB property of water has often been cited as the most important feature of the molecule. The high density of HBs per unit volume makes water a particularly good solvent for polar material and also contributes to its ability to ionize solutes. Furthermore, the strength of these bonds, about 10 kT_{room}, is why water is liquid at room temperature when other simple liquids are not. Despite these large energies, HBs in ambient water are broken and formed again frequently. Experimental measurements have shown that individual waters reorient themselves every 2 ps and move the distance of one molecule diameter every 7 ps.

4.1.2.5 Tetrahedral Structure

The four-fold nature of water hydrogen bonding is what leads to the tetrahedral structure of water networks in both the liquid and solid states. This tetrahedral structure is rigidly maintained in all known forms of ice. As ice melts, waters deviate from this four fold coordination as the angles and distances of the HBs become more relaxed; larger and smaller ring structures can form and interstitial spaces can be reduced, leading to the increased density. With increasing temperature, bond lengths become longer and the volume of the liquid begins to increase. Taking together, these properties explain the maximum density of 4 °C for H_2O, 11.2 °C for D_2O and 13.4 °C for T_2O.

One should note that the electrostatic potential of water is not tetrahedral. Rather the four-fold hydrogen bonding where the lone pairs collectively accommodate two hydrogen bonds leads to this structure. This has become apparent in molecular models that have attempted to exploit a tetrahedral charge distribution. In these cases, the water networks have been overly structured.

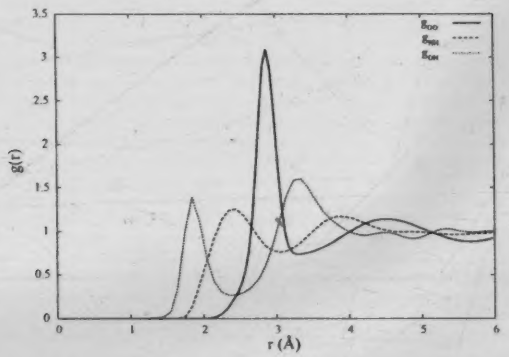


Figure 4.4: Radial distribution functions of water. Plotted using data from [71].

4.1.2.6 Intermolecular Structure

Tetrahedral coordination in water leads to its characteristic pair correlation function (Fig. 4.4). In particular, the peaks in this function correspond to HB lengths in first and second (or large) neighbours. As the temperature increases there is an increased spread in HB lengths and angles, resulting in a softening of the peaks.

4.1.2.7 Proton Conduction

Proton conduction is still not completely understood. However, its appears to be the results of hydrogen bonding and polarizability of the water molecule. At least one possible scenario (perhaps the oldest) is the Grotthuss mechanism [63, 72] where a free hydrogen associating with a lone pair of a water oxygen. The oxygen can then exchange one of its hydrogens for the formerly free hydrogen. This process continues as a chain reaction, forming so-called water-wires.

4.2 Proteins-Water Interactions

Proteins have evolved in an aqueous environment; without this environment, they cease to function. Regardless of specific details for individual proteins, the dominant mechanism for water-protein interactions is hydrogen bonding, or the lack thereof. This is seen clearly in both protein folding and protein-protein binding and is explored in Sections 4.2.1 and 4.2.2.

To be clear, a substance can be hydrophilic, hydrophobic or neither [73]. A substance is hydrophilic if it is soluble in water. If it is not soluble in water but is soluble in a non-aqueous solvent (e.g. hydrocarbons) then it is hydrophobic.

4.2.1 Protein Folding and Structure

Broadly speaking, folding is the result of different regions of the protein being hydrophilic - forming hydrogen bonds with water - and hydrophobic - not forming hydrogen bonds with water. We begin our discussion with the parts of the protein that don't interact favourably with water and then those that do.

4.2.1.1 Hydrophobicity

The most basic principle in protein folding is that hydrophobic side chains are packed into the center of the protein while hydrophilic ones are left to interact with water [73, 74]. Qualitatively, this has been known for a considerable time. However, to make predictions about protein structure or stability, there must be a quantitative understanding as well. Two major advances in this regard have been the Kauzmann and the solvent accessible surface area (SASA) models [73].

Noting that the interior of most proteins was hydrophobic, Kauzmann suggested that this could be modelled as an oily liquid. The free energy change of burying a side chain in the interior of the protein could be estimated experimentally with a model compound for the side-chain and an oily liquid to model the protein interior. The model compound could then be partitioned between water and a non-aqueous solvent to obtain the solvation free energy. For example, toluene was used as a model for the phenylalanine side-chain with the free energy change given as

$$\Delta G_{\text{hyd}}^{\text{toluene}} = -RT \ln x = -5.45 \text{ kcal/mol} \tag{4.1}$$

where x is the mole fraction of toluene in the water. Early work done by Kauzmann and others in this area indicated that hydrophobicity is the largest energetic factor in protein folding.

The method of using side-chain analogous is still used in both computational and experimental studies for predicting the free energy of solvation [75, 76]. However, partitioning is typically between water and a vacuum instead of a hydrophobic solvent. This is, at least in part, because it is not clear what the hydrophobic liquid should be. In fact, there may not be a suitable liquid as the core of typical proteins is densely packed, approaching the optimal densities found for hexagonal close packed or face centered cubic lattices.

In the 1970s it was proposed that there was a linear relationship between the SASA and the transfer free energy

$$\Delta G_{\text{hyd}} = k_{\text{h}} (\text{SASA}_{\text{F}} - \text{SASA}_{\text{U}}) \tag{4.2}$$

where F and U are the folded and unfolded states and k_{h} is an empirically determined proportionality constant. This has been largely based on experimental data for the solvation free energy of linear alkanes that show a linear relation between the SASA and ΔG_{hyd}. However, when branched or cyclic hydrocarbons are used, the result is a non-linear correlation. Due to the computational simplicity, the model is still extensively used (see Section 4.3.2.3) but is generally held to be a rough approximation at best.

The true free energy cost of hydration consists of the entropy and enthalpy change as associated with taking the solute from a vacuum into the solvent

$$\begin{aligned} \Delta G_{\text{hyd}} &= \Delta H_{\text{hyd}} - T\Delta S_{\text{hyd}} \\ &= \Delta U_{\text{hyd}} + P\Delta V - T\Delta S_{\text{hyd}} \end{aligned} \tag{4.3}$$

where we have split the enthalpy into energetics (U) and cavity formation in the solvent ($P\Delta V$). However, approaches that exploit this relation, such as thermodynamic integration, tend to be computationally expensive and full solvation calculations are limited to small molecules of a few tens of atoms at most [77]. This makes implicit solvents and SASA methods relatively popular.

The relative contributions of each term in Equation (4.3) are known to vary with temperature. In particular, $S_{\text{hydration}}$ has been observed to decrease linearly with increasing temperature [73]. For both protein unfolding and hydrocarbons in aqueous solution, $S_{\text{hydration}}$ goes to 0 at approximately 110 °C. The reduction of the entropic barrier goes far to explaining heat denaturing for proteins. Cold denaturing is caused by quite different circumstances (see Section 4.2.1.3).

While the role of hydrophobic residues in protein folding has been well established, the physical details of how these residues go from an aqueous environment to a dry one are still unknown. The two competing theories for the process are the dewetting effect and expulsion mechanism [74].

The dewetting effect postulates that there is a decrease of water density around hydrophic residues followed by a spontaneous collapse of the core or the protein to the folded state. The main evidence in support of this theory has been dewetting experiments and simulations of idealized parallel hydrophobic plates [78]. These show a vapor layer forming around large hydrophobic surfaces. In the case of these idealized situations, a vacuum is formed and the plates quickly collapse together. However, this has not been observed in experiment or simulation for real proteins.

In the expulsion mechanism theory, hydrophobic residues remain hydrated throughout the folding process and are only expelled slowly as the protein arrives at its final conformation. In fact, the water is necessary as a lubricant for the folding process. This is characterized by a slow folding process with several intermediates rather than the fast two-state process for dewetting.

4.2.1.2 Hydrogen Bonding

The hydrophobic residues' interactions with water are dominated by the second and third terms of Equation (4.3). There is little energetic contribution, other than steric, to this equation as these non-polar solutes do little to polarize the solvent. Thus, larger solutes are more hydrophobic (ΔV) as are ones that tend to locally order water. However, all residues have polar backbone atoms, NH and CO, that can form hydrogen bonds with water and other residues (backbone or side-chain). The backbone can, in theory, form three hydrogen bonds; one hydrogen is donated by NH while CO can accept two. Internal hydrogen bonding in the form of β-sheets and α-helices also serves to satisfy the polar groups of the backbone that are buried inside the otherwise hydrophobic core or the protein.

Hydrophilic residues, on the other hand, still incur the cost of the cavity formation and entropic terms but compensate by forming hydrogen bonds with water. There are at least two roles for hydrophilic residues in protein folding. By favourably interacting with the solvent, they help contribute to the hydrophobic folding mechanism. The other key contribution is making the protein soluble in water. For example, ALA_{13} is not soluble in water unless capped by two basic amino acids [73].

ALA_{13} capped by basic residues is also interesting because it is helix forming. Until this experiment, it was believed that all helices needed to be stabilized by salt bridges or tertiary structure. Experimental measures of helix formation show, on average, that enthalpy favourably accounts for half of the free energy of folding this helix.

4.2.1.3 Protein Induced Water Structure

As water shapes the native structure of proteins, proteins induce structure in water. Water has different properties depending on which of four different parts of this structure it resides in: buried, first hydration shell, second hydration shell or bulk [74, 78–81].

Buried Buried waters interact the most strongly with the protein; they are a part of the proteins structure. Most commonly, they directly hydrogen bond to the backbone, stabilizing irregular structures [78] and are characterized by long residence times (> 1 ns) [74, 79, 80].

First Hydration Shell The first hydration shell includes waters that are not buried but are in direct contact with both the protein and the bulk water. The density of this water is 10-20% higher than bulk [74, 79] with longer residence times as well (500 ps) [74]. About 55% of this water binds directly to the backbone [74]. The structure has been compared to supercooled water at 268 K and generally does not freeze with the bulk solvent [81]. Furthermore, this layer has been shown to be necessary and, often, sufficient for protein activity [81].

While this primary shell may not freeze completely, its order and structure have a profound effect on protein structure. Most proteins have a cold denaturation threshold, often near or below the freezing temperature of bulk water [74, 82]. While the exact mechanism is unknown, MD studies have identified that upon cooling, water associates more with hydrophobic groups. One hypothesis is that the water becomes more structured by the bulk and the strength and number of water-protein hydrogen bonds is reduced. Thus, the energetic difference between hydrophobic and hydrophilic interactions is reduced, and a compact, folded state become less favourable.

Second Hydration Shell A second hydration shell can often be found 5-7 Å from the surface of the protein [79]. The density increase is much small than for the first hydration shell and diffusion rates are intermediate to those of the first hydration shell and the bulk. However, there is a strong anisotropy with perpendicular rates being faster than parallel rates.

Bulk As the name would suggest, these waters have the same physical properties as bulk water. Typically, these waters have residence times of 5-7 ps and must be out of the range of the proteins screened electrostatic potential, which can be 15 Å or more from the protein [79]. There is, however, a smooth transition from waters in the second hydration shell to the bulk.

4.2.2 Protein Interactions

The same physical mechanisms of protein hydration at play in protein folding are also at play in protein-protein binding (or binding in general). The major difference is that binding sites must alternate between wet and dewetted states. Since this requires the removal of considerable amounts of water from both interfaces, water plays a significant role in binding.

Binding interfaces can be characterized by three main groups, depending on their level of hydration [74, 78]: sites that have no waters when bound (typically co-folded), interfaces with multiple dry binding sites surrounded by water and partially hydrated binding sites. The direct role of water is two fold. The first role is to mediate and smooth the long range interactions, allowing the reactants to find the optimal configuration without getting stuck in an intermediate state. Once the correct configuration has been established, water may be part of the final structure, bridging otherwise disconnected residues via hydrogen bonding. The later role, in particular, is thought to allow promiscuous binding (i.e. a single binding site can have multiple targets) which may be necessary for signaling pathways or transient interactions.

4.3 Molecular Modelling of Solvent

In the context of molecular modelling there are two basic approaches to solvation, explicit and implicit. Explicit solvent models explicitly represent all of the molecules (and typically atoms) in the system. These models treat the solvent molecules in the same manner as the solute. Implicit models treat the solvent as a continuum without discrete particles. Often the solvent is a featureless bulk dielectric medium. Occasionally, the methods are combined, with an explicit solvent near the surface of the solute and implicit solvent beyond a given radius.

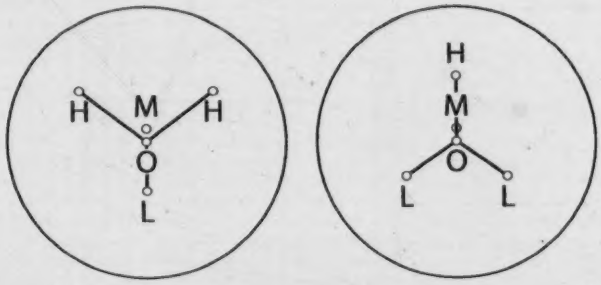


Figure 4.5: Schematic representation of the SPC and TIP families of water models shown in two orientations. The point particles are oxygen (O), hydrogens (H), the ∠HOH bisector (M) and lone-pairs (L). Only oxygen and hydrogens have mass. Only oxygen has a van der Waals radius.

Of course, solvents include, but are not limited to, water. Water is the most common solvent for biological materials and the only one used in this thesis. Unless otherwise noted, details of specific models deal with water though the general approach is applicable to all solvents.

4.3.1 Explicit Water Models

A wide variety of explicit water models have been developed, including 2-D (e.g. Mercedes Benz model [83]), polarizable (e.g. Yu *et al.* [84]) and quantum variants (e.g. Lobaugh *et al.* [85] and Bernal-Uruchurtu *et al.* [86]). For a historical perspective of explicit water model development see Finney [64] or Guillot [65].

The vast majority of molecular simulations use the TIP3P [69] or SPC/E [87] water models. Thus we will briefly review these models. It is important to keep in mind that these are fixed charge models with no temperature dependence for any of the parameters. Though some have been parameterized to reproduce water properties over a wide range of temperatures, most have been optimized for ambient temperature and pressures and should be used only at these values.

For an in-depth comparison of these and other models see Guillot [65]. Vega *et al.* [88] complements this with an overview of the most commonly used models and their melting temperatures.

4.3.1.1 TIP3P and Family

The transferable intermolecular potential functions (TIPS) is the basis of all of the TIP models, produced by the Jorgensen group, and derivatives there of. While originally developed for water, alcohols and ethers, TIP is synonymous with water due to the popularity of the TIP3P model. TIP3P is include as the default water model for Amber and CHARMM molecular modelling packages.

TIP TIP [89] was the original model published in 1981. The motivation was to create a simple, modular potential and parameters that could be applied to a wide variety of molecules. The form of the TIP potential is

$$\Delta E = \sum_a \sum_b \left(\frac{q_a q_b e^2}{r_{ab}} + \frac{A_a B_b}{r_{ab}^{12}} + \frac{C_a C_b}{r_{ab}^{6}} \right) \tag{4.4}$$

Model	q_O	q_H	q_M	q_L	ϵ_O	σ_O	r(OH)	r(OM)	r(OL)	∠HOH	∠LOL
	e				kcal/mol	Å	Å			°	
TIP	-0.8000	0.40000	-	-	0.11880	3.6090	0.9572	-	-	104.52	-
TIP3P	-0.8340	0.41700	-	-	0.15200	3.5534	0.9572	-	-	104.52	-
TIP4P	0	0.52000	-1.04000	-	0.15500	3.5398	0.9572	0.150	-	104.52	-
TIP4P-Ew	0	0.52422	-1.04844	-	0.16275	3.5519	0.9572	0.125	-	104.52	-
TIP5P	0	0.24100	-	-0.482	0.16000	3.5020	0.9572	-	0.7	104.52	109.47
SPC	-0.8200	0.41000	-	-	0.15530	3.5534	1.0000	-	-	109.28	-
SPC/E	-0.8476	0.42380	-	-	0.15530	3.5534	1.0000	-	-	109.28	-

Table 4.2: TIP and SPC family water model parameters.

Here, r is the distance between atoms a and b, the respective charges are give by q and the last two terms of the equation are a 6-12 Lennard-Jones potential with A and C as parameters. Using the Amber/CHARMM form of the Lennard-Jones potential, it has the form

$$\Delta E = \sum_a \sum_b \left(\frac{q_a q_b e^2}{r_{ab}} + \epsilon \left(\frac{\sigma^{12}}{r_{ab}^{12}} + 2\frac{\sigma^6}{r_{ab}^6} \right) \right) \tag{4.5}$$

where ϵ is the well depth and σ the radius. Parameters for this model can be found in Table 4.2. The bond length and angle parameters were taken from experimental measurements of water in the gas phase. Also, note that only oxygen has a Lennard-Jones potential. See Figure 4.5.

This model is not in current use; it is included as it is the origin of the TIP based water model.

TIP3P TIP3P [69] was a re-parameterization of the original TIPS model using Equation (4.4) with the same geometry (see Figure 4.5). While the calculated energy and density were improved it was found that this was at the cost of the second g_{OO} peak.

TIP4P TIP4P [69] was created to overcome the problems found with the TIP3P model and was one of two parameter sets for a four-site model originally proposed by Bernal and Fowler in 1933 [90]. In both the TIP4P and TIPS2 (introduced in the same paper) models the negative charge is removed from the oxygen and placed along the ∠HOH bisector at the center-of-mass of the molecule (see Figure 4.5). TIP4P was, however, modified to reproduce the density of liquid water at 25°C and 1 atm.

TIP4P-Ew TIP4P-Ew [91] was a re-parameterization of the TIP4P model intended to reproduce the properties of liquid water at 1 atm over typical physiological temperature range using Ewald summation. All of the Jorgensen group's models had been produced for the use of a cutoff in calculating electrostatic interactions. Using Ewald summation introduces a systematic decrease in the pressure and increase in the diffusion constant for all of the previous models.

Notably, the model is able to reproduce many temperature dependent properties of water without an explicit temperature dependence in the model. In particular, it was parameterized by fitting experimental data over the range of 235.5-400 K for the bulk density and enthalpy of vaporization. It generally outperforms other models for these values. Other values, such as the self-diffusion coefficient are also predicted in agreement with experiment. Other values, show systematic errors, such as the dielectric constant which is consistently low.

TIP5P TIP5P [92] created to predict the temperature and pressure dependent properties of water over the ranges -37 to 100°C and 1 to 10000 atm. In particular, the temperature dependence of density was considered a key goal.

As four-site models were not found to be fruitful, they adopted a five site model with a tetrahedral structure. The model has an oxygen with a mass and Lennard-Jones parameters but no charge. There are four charge sites corresponding to two hydrogen and two lone-pair sites. The geometry remains the same as the original TIP model but with addition of the lone-pair sites. See Figure 4.5.

4.3.1.2 SPC and Family

SPC water and family have been the main alternative to the TIP family in biological molecular simulations. The potential shares the same form as Equation (4.5) but has a modified geometry and different parameterization.

SPC The simple point charge (SPC) model [87] was developed to provide a fast, accurate effective pair potential model for molecular simulations of proteins. It also uses Equation (4.5) for its potential. The geometry, however, is slightly different than that of the TIP family, closer to the values for liquid water. After setting the geometry and assigning a charge of q to the hydrogens and $-2q$ to the oxygen, three parameters where left to fit the model, q and the Lennard-Jones parameters for oxygen. The attractive r^6 term was taken from the experimental value from the London expression. The resulting 2D parameter space was searched with 12 NVT MD simulations with a density of 1 g/cm^3 and temperature of 300 K. The resulting parameters can be found in Table 4.2.

SPC/E To improve the quality of the SPC model and demonstrate the importance of the polarization, the extended SPC (SPC/E) model was created [70]. This time, four models were produced that modified the charges of the original SPC model and created a larger dipole. NPT MD simulations were run at 300K and 1 atm. The model that best fit the experimental values for potential energy, self diffusion constant and density was selected.

4.3.2 Implicit Solvents

Implicit solvents, generally speaking, ignore the molecular nature of the solvent, treating it as a bulk. As a result, these models generally do not reproduce the effects of solvent as well as explicit models. However, they are generally faster to calculate, reduce the overall degrees of freedom of the system and enhance sampling. As the accuracy of these models has improved, so has their popularity.

A wide variety of approaches has been taken to implicit solvation. The simplest forms have been constant or distance dependent dielectric coefficients. The related generalized Born (GB) and Poisson-Boltzmann (PB) models are generally recognized as the most successful and are the most widely used. However, these only account for polar solvation effects. Apolar contributions are usually calculated with a simple solvent accessible surface area term.

4.3.2.1 Poisson-Boltzmann

PB has become the gold standard in implicit solvent calculations for molecular modelling [93, 94]. This is, in large part, due to its firm theoretical foundation. However, as this is a relatively expensive model to calculate, it has not found much use in MD calculations. Rather, it has

been used in MD in the course of researching implicit solvation methods or for calculating single point energies, for example Baker *et al.* [95].

The theoretical base is the Poisson equation

$$-\nabla \cdot \epsilon(x)\nabla\psi(x) = 4\pi e^2 \rho(x) \tag{4.6}$$

where e is the elementary charge. This assumes a fixed charge distribution, $\rho(x)$, to calculate the potential, $\psi(x)$. The position dependent dielectric coefficient, $\epsilon(x)$, is generally taken to be 2 inside the protein and 80 outside if the solvent is water.

As physiologically or experimentally relevant solvents also contain free ions, this is added into the theory via a Boltzmann distribution

$$\nabla \cdot [\epsilon(x)\nabla\psi(x)] = -4\pi e^2 \rho(x) - \sum_{i=1}^{m} q_i c_i e^{(-q_i\psi(x) - V_i(x))/kT} \tag{4.7}$$

The additional term includes m species of ion with concentration c_i and charge q_i. The steric interaction with the fixed solute is given in $V_i(x)$.

In practice, we have equal concentrations of two oppositely charged monovalent ions, which means $(-q\psi(x) - V(x)) \ll kT$ and $q = q_1 = q_2$, $c = c_1 = c_2$. Thus, we use the following approximation for the Boltzmann term [96]

$$\begin{aligned} \sum_{i=1}^{2} qce^{(-q\psi(x) - V(x))/kT} &= qc\sinh((-q\psi(x) - V(x))/kT) \\ &\approx qc\,(-q\psi(x) - V(x))/kT \end{aligned} \tag{4.8}$$

This gives the linearized form of Equation (4.7)

$$\nabla \cdot [\epsilon(x)\nabla\psi(x)] = -4\pi e^2 \rho(x) + qc\frac{q\psi(x) + V(x)}{kT} \tag{4.9}$$

Equation (4.9) must still be solved numerically. Though other methods exist, the multi-grid, finite difference approach has been the most economical and popular. The solution is calculated on a 3D grid with a typical grid spacing of 0.25-0.5 Å. Properly implemented, it can also be efficiently parallelized [95].

To find a solution a set of boundary conditions must also be applied. Most commonly the potential is calculated at the boundary using a multipole expansion of the solute charge distribution, which is truncated after the first few terms. Often a technique, known as focusing, is used achieve high resolution of a small part of a system. First, the potential is calculated on a large, coarse grid. From this coarse grid the boundary conditions for the region of interest are determined.

4.3.2.2 Generalized-Born

Like PB, GB is a macroscopic, continuum model for solvation. However, GB avoids cost of solving the PB equation with a simple analytical approximation. First, the total electrostatic free energy of widely separated particles of van der Waals radius ρ in a dielectric medium, with coefficient ϵ, is given as

$$\Delta G_{\text{elec}} = 332 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{q_i q_j}{r_{ij}\epsilon} - 166\left(1 - \frac{1}{\epsilon}\right) \sum_{i}^{n} \frac{q_i^2}{\rho_i} \tag{4.10}$$

where the second term is the Born equation [97]. The first term can be rewritten to give Coulomb's law *in vacuo* and a corrective term

$$\Delta G_{\text{elec}} = 332 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{q_i q_j}{r_{ij}} - 332 \left(1 - \frac{1}{\epsilon}\right) \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{q_i q_j}{r_{ij}} - 166 \left(1 - \frac{1}{\epsilon}\right) \sum_{i}^{n} \frac{q_i^2}{\rho_i}$$

$$= 332 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{q_i q_j}{r_{ij}} - 166 \left(1 - \frac{1}{\epsilon}\right) \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{q_i q_j}{f_{\text{GB}}} \tag{4.11}$$

f_{GB} is a function of r_{ij} and α_{ij} that is smooth and not uniquely defined. For isolated particles, α_{ij} will equal their van der Waals radii. However, if a particle is buried inside of a larger molecule, α will be the effective Born radius and may be approximated by the radius of the enclosing molecule. The function of f_{GB} is to smoothly interpolate between the case when the two spheres are merged (no dielectric screening, $r_{ij} \to 0$) and when the particles are very far apart (point particles in a dielectric, $r_{ij} \to \infty$). The most common form is

$$f_{\text{GB}} = \sqrt{r_{ij}^2 + \alpha_i \alpha_j e^{-r_{ij}^2/4\alpha_i\alpha_j}} \tag{4.12}$$

Calculating accurate effective Born radii is critical to the success of a GB implementation. As they must be recalculated every time the conformation changes, it is also essential that the method be computationally efficient. A typical expression for the effective radius is the Coulomb field approximation and is used in Amber [98]

$$\alpha_i^{-1} = \rho_i^{-1} - \frac{1}{4\pi} \int_{\text{solute}} \frac{\theta(|\mathbf{r}| - \rho_i)}{r^4} d^3\mathbf{r} \tag{4.13}$$

where θ is the Heaviside function. The volume integral of this function is, in turn, approximated to create a closed-form analytical expression. The simplest of these expressions is to integrate over the van der Waals radii of the particles. Since this leads to gaps in the interior of large molecules where water would not actually be able to fit, significant errors can arise from this method. More sophisticated approximations have been created with the intent of limiting such artifacts and are the currently preferred methods [98].

It is possible to calculate so-called perfect Born radii by solving the PB equation but this defeats the purpose of GB. This has been done to investigate the limits of GB and has been shown that even with perfect radii, GB is not as accurate as PB [77].

Ion Screening To accurately simulate *in vivo* and *in vitro* conditions it is important to include salt ions. The approach used for GB starts with the solution to the linearized PB equation for two widely separated ions [99]

$$\Delta G_{pol} = -332 \left(1 - \frac{e^{-\kappa r_{ij}}}{\epsilon}\right) \frac{q_i q_j}{r_{ij}} \tag{4.14}$$

where κ is the Debye-Hückel screening parameter. For the case of a single ion in solution we have

$$\Delta G_{pol} = -116 \left(1 - \frac{1}{\epsilon}\right) \frac{q_i^2}{\rho} - \frac{q_i^2 \kappa}{2\epsilon(1 + \kappa a)} \tag{4.15}$$

a is the distance to which ions are excluded so $a - \rho$ is the ion exclusion radius. To a close approximation, both these limits are satisfied with the substitution

$$\left(1 - \frac{1}{\epsilon}\right) \to \left(1 - \frac{e^{-\kappa f_{\text{GB}}}}{\epsilon}\right) \tag{4.16}$$

giving the final equation

$$\Delta G_{\mathrm{GB}} = -166\left(1 - \frac{e^{-\kappa f_{\mathrm{GB}}}}{\epsilon}\right)\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{q_i q_j}{f_{\mathrm{GB}}} \tag{4.17}$$

4.3.2.3 Solvent Accessible Surface Area

The solvation free energy can be decomposed into three terms [97]

$$\Delta G_{\mathrm{sol}} = \Delta G_{\mathrm{cav}} + \Delta G_{\mathrm{vdW}} + \Delta G_{\mathrm{pol}} \tag{4.18}$$

where G_{cav} and G_{vdW} are the cost of creating a cavity in the solvent and the van der Waals interaction between the solute and solvent. G_{pol} is the solute-solvent electrostatic polarization term and is calculated through PB or GB.

While more sophisticated approaches do exist [94], the nonpolar terms are commonly modelled using an SASA model (see Section 4.2.1.1)

$$\Delta G_{\mathrm{nonpol}} = \Delta G_{\mathrm{cav}} + \Delta G_{\mathrm{vdW}} = \sum_k \sigma_k \mathrm{SA}_k \tag{4.19}$$

SA_k is the solvent accessible surface area (SASA) of the atom type k and σ_k is the empirical solvation parameter for this atom type. The physical justification for this is the observation for fully saturated hydrocarbons, G_{sol} is linearly related to the SASA and G_{pol} should be zero. In practice, however, there is no distinction made between atom types when the method is used in simulations and a single σ is multiplied by the total SASA of the solute. In addition, the value of σ can vary considerably from 0.005-0.070 kcal/mol [100] though 0.005 kcal/mol is the most common choice.

Chapter 5

Molecular Dynamics Simulations

Molecular Dynamics (MD), on its most basic level, is the brute force integration of Newton's equations of motions for a large number of atoms interacting via a given potential energy function. This involves calculating the force on every atom imparted by every other atom in the system. Unlike other numerical methods, such as Metropolis Monte Carlo methods, in MD we not only collect information about the equilibrium properties but, also, the time dependence, allowing the calculation of transport properties as well.

We must keep in mind that this is a classical treatment of systems that are bordering on the quantum world. While a good classical potential can approximate most physical properties, such as chemical bonds (itself only a model) and electrostatics, reasonably well, not all properties of the system can be captured. This is in large part due to the limits of real world computational power. However, as computers increase in speed so too does the quality of potentials that may be used.

As the field has matured, the sophistication of the systems being studied and the programs being used to model them has grown. Thus, unless the model being used is quite simple, it is standard practice to use one of several software packages that have been developed over the past 20 years. Some of the most popular packages over the years have been Amber [5], CHARMM [101, 102], GROMACS [103] and NAMD [104]. All of these come with their own potentials and many other popular potentials and software packages exist as well.

Far too many algorithms have been developed for MD to discuss them all in this chapter. Rather, we focus on the main algorithms used in the remaining chapters of this thesis and discuss how they relate to alternative methods. Sections 5.1 to 5.4 introduce algorithms intended to reproduce accurate physical and physiological conditions in our systems. Topics covered, in order, are integration of the equations of motion, reproduction of rigorous ensembles, minimization of the effects of boundary conditions and potential energy functions governing atom-atom interactions. The final two sections introduce methods to increase computational efficiency. First, methods are considered to reduce the cost of integrating the equations of motion, such as distance based cutoffs. The chapter concludes with a discussion of advanced sampling techniques, such as replica exchange and adaptive biasing force MD.

5.1 Equations of Motion

As stated above, MD is the integration of Newton's equations of motion of a molecule using a classic Hamiltonian to derive forces. As it is not possible to do this analytically we must develop numerical methods for doing so. The method that we choose should conserve the total

energy of the system and be time reversible, just as Newton's equations are. We would like the algorithm to be fast but, since most computation time will be spent evaluating the forces on the particles, this is not the most important consideration. The simplest method (and in many cases one of the best) that does all this for us is the Verlet method [105]. Other popular methods include the Velocity Verlet method, the leap-frog algorithm and the Multi Time Step (MTS) method.

5.1.1 Verlet

To derive the Verlet algorithm we start by Taylor expanding the position $\mathbf{r}$ one time step ahead about the time t [106].

$$\mathbf{r}(t+\Delta t) = \mathbf{r}(t) + v(t)\Delta t + \frac{\mathbf{f}(t)}{2m}\Delta t^2 + \frac{\Delta t^3}{3!}\mathbf{r}''' + \mathcal{O}(\Delta t^4) \tag{5.1}$$

$$\mathbf{r}(t-\Delta t) = \mathbf{r}(t) - v(t)\Delta t + \frac{\mathbf{f}(t)}{2m}\Delta t^2 - \frac{\Delta t^3}{3!}\mathbf{r}''' + \mathcal{O}(\Delta t^4) \tag{5.2}$$

$$\mathbf{r}(t+\Delta t) + \mathbf{r}(t-\Delta t) = 2\mathbf{r}(t) + \frac{\mathbf{f}(t)}{m}\Delta t^2 + \mathcal{O}(\Delta t^4)$$

$$\mathbf{r}(t+\Delta t) \approx 2\mathbf{r}(t) - \mathbf{r}(t-\Delta t) + \frac{\mathbf{f}(t)}{m}\Delta t^2 \tag{5.3}$$

The Verlet algorithm is accurate to order Δt^4. Although the velocity is not used in the Verlet scheme it is useful to calculate it in order to compute the kinetic energy of the system. The equation for velocity can be derived in a similar way to the position as we have seen above

$$\mathbf{r}(t+\Delta t) - \mathbf{r}(t-\Delta t) = 2\mathbf{v}(t)\Delta t + \mathcal{O}(\Delta t^3)$$

$$\mathbf{v}(t) = \frac{\mathbf{r}(t+\Delta t) - \mathbf{r}(t-\Delta t)}{2\Delta t} + \mathcal{O}(\Delta t^2) \tag{5.4}$$

This is only accurate to order δt^2.

While still a good algorithm the Verlet algorithm is not often used. This is primarily due to the fact that the $\mathcal{O}(\Delta t^4)$ accuracy of the positions is poor compared to some other methods. The lack of accuracy impacts the length of the time step we are able to employ which, in the interest of computational efficiency, we would like to be as long as possible. The $\mathcal{O}(\Delta t^2)$ precision of the velocities is even worse, though this is only used for calculating the temperature of the system. Despite its short comings, the Verlet method is time-reversible (which cannot be said about many more accurate algorithms) and is not susceptible to long-time energy drift.

Three widely used alternatives to the Verlet method are the velocity Verlet, leapfrog Verlet and Beeman methods. Of the three the Beeman method is the only one that differs significantly from the basic Verlet method. It can be shown that the positions produced by the Beeman method satisfy the Verlet algorithm and the velocities are more accurate. Unfortunately, the scheme is not time-reversible[106].

5.1.2 Leapfrog Verlet

The leapfrog Verlet method offers improved accuracy over the basic Verlet method and is the default method in CHARMM. The algorithm evaluates the velocities at half-integer time steps and uses the velocities to calculate the new positions[106]. Because the velocities and positions are not defined at the same time the total energy of the system cannot be directly calculated for the system. Velocities at time t are typically calculated as the average of $\mathbf{v}(t-\Delta t)$ and $\mathbf{v}(t+\Delta t)$.

To derive the leapfrog method we begin by rewriting Equation (5.4)

$$\mathbf{v}(t) = \frac{\mathbf{r}(t + \Delta t/2) - \mathbf{r}(t - \Delta t/2)}{\Delta t} \tag{5.5}$$

where we have simply substituted $\Delta t/2$ for $\Delta t'$. Note that this makes the velocity slightly more accurate. Having done this we can state the velocities at half integer time steps about the time of interest

$$\mathbf{v}(t - \Delta t/2) = \frac{\mathbf{r}(t) - \mathbf{r}(t - \Delta t)}{\Delta t} \tag{5.6}$$

$$\mathbf{v}(t + \Delta t/2) = \frac{\mathbf{r}(t + \Delta t) - \mathbf{r}(t)}{\Delta t}. \tag{5.7}$$

From Equations 5.6 and 5.7 we immediately obtain the previous and new positions respectively

$$\mathbf{r}(t - \Delta t) = \mathbf{r}(t) - \Delta t \mathbf{v}(t - \Delta t/2) \tag{5.8}$$

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \Delta t \mathbf{v}(t + \Delta t/2). \tag{5.9}$$

Substituting these equations into Equation 5.3 we have

$$\mathbf{r}(t) + \Delta t \mathbf{v}(t + \Delta t/2) = 2\mathbf{r}(t) - \mathbf{r}(t) + \Delta t \mathbf{v}(t - \Delta t/2) + \frac{\mathbf{f}(t)}{m}\Delta t^2$$

$$\mathbf{v}(t + \Delta t/2) = \mathbf{v}(t - \Delta t/2) + \frac{\mathbf{f}(t)}{m}\Delta t \tag{5.10}$$

Due to using the more accurate velocities at half integer time steps the leapfrog Verlet method is superior to the basic Verlet method even though the two yield identical trajectories.

5.1.3 Velocity Verlet

The velocity Verlet method provides the additional accuracy of the leapfrog method with velocities and positions computed at equal times[106]. We begin with the Taylor expansion forward and backward in time (Equations 5.1 and 5.2 respectively). For the new position we merely truncate Equation 5.1 beyond the Δt^2 term.

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \mathbf{v}(t)\Delta t + \frac{\mathbf{f}(t)}{2m}\Delta t^2 \tag{5.11}$$

One should note that this equation, by itself, is the Euler algorithm. The Euler algorithm is not time-reversible or area preserving. Furthermore, it suffers from a massive energy drift that causes the integration to become unstable and eventually "blow-up". The velocity Verlet algorithm avoids this by using a different velocity update scheme. To obtain this update scheme we take Equation 5.2 (truncated beyond Δt^2 and evaluate it at $t' = t + \Delta t$.

$$\mathbf{r} = \mathbf{r}(t + \Delta t) - \mathbf{v}(t + \Delta t)\Delta t + \frac{\mathbf{f}(t + \Delta t)}{2m}\Delta t^2 \tag{5.12}$$

Substituting this into Equation 5.11 and solving for $\mathbf{v}(t + \Delta t)$ we obtain

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t + \Delta t) - \mathbf{v}(t + \Delta t)\Delta t + \frac{\mathbf{f}(t + \Delta t)}{2m}\Delta t^2 + \mathbf{v}(t)\Delta t + \frac{\mathbf{f}(t)}{2m}\Delta t^2$$

$$\mathbf{v}(t + \Delta t) = \mathbf{v}(t) + \frac{\mathbf{f}(t + \Delta t) + \mathbf{f}(t)}{2m}\Delta t \tag{5.13}$$

Like the leapfrog method the velocity Verlet method produces identical trajectories to the standard Verlet method but with increased precision.

5.1.4 Langevin Dynamics

An alternative to Newtonian dynamics is Langevin dynamics (LD). It is generally used to simulate particles immersed in a solvent or in contact with a heat bath.

When used as a heat bath LD may be used with implicit or explicit solvent, though the parameters used will vary. In the simplest form of implicit solvation, a uniform bulk dielectric coefficient (e.g. $\epsilon = 80$) is typically used. Since this method eliminates the need for water molecules, it provides a considerable speedup compared to a system fully hydrated with explicit water.

Unfortunately, this bulk approach to system hydration has many drawbacks. Among them is the fact that all the atoms are subjected to the same LD and ϵ even if they are deeply buried inside the protein. Thus, a normally hydrophobic residue would be subjected to the friction and random forces of the bulk solvent. Furthermore, this residue would normally see its electrostatic neighbours through a dielectric constant of 1 but the imposed bulk dielectric constant greatly changes the observed potential.

Needless to say, this is no longer common practice for modern simulations. When implicit solvent is used it is typically generalized-Born [107], Poisson-Boltzmann [95] or 3D-RISM [108]. In these cases, LD may be considered MD modified by addition of a friction coefficient controlling the coupling to a heat bath. It is very efficient for this purpose and eliminates some of the other problems with other heat baths. Notably, the formation of hot spots or 'flying ice-cubes' is not a problem with LD. The cost of this is that velocity information is lost due to the stochastic nature of the equation.

The Langevin equation for a single particle has the form [109–111]

$$m\frac{d\mathbf{v}}{dt} = -\gamma m\mathbf{v} + \mathbf{F}(t) + \mathbf{A}(t) \tag{5.14}$$

where $\gamma = 6\pi a\eta/m$ is the Stokes law friction constant and $\mathbf{A}(t)$ is a stochastic force imparted by the solvent. The stochastic force term has the form of a Gaussian with a mean of 0 and an intensity of

$$\langle \mathbf{A}(t)\mathbf{A}(t+\Delta t)\rangle = 6\gamma k_B T\delta(\Delta t). \tag{5.15}$$

Typically, the Langevin equation is used in one of three regions [111]. In the case $\gamma\Delta t \ll 1$ the solvent does not greatly activate or deactivate the solute since the timescale is short compared to the velocity relaxation time. If $\gamma\Delta t >> 1$ then the velocity relaxation is shorter than the timescale and the motion is greatly damped. The third region is then in between these two extremes. The Langevin integrator used in CHARMM, NAMD, Amber and GROMACS was derived for the first case, $\gamma\Delta t \ll 1$.

The derivation of the discrete form of Equation (5.14) that is stable for $\gamma\Delta t \ll 1$ begins with the simple finite difference scheme [111]

$$\frac{d^2\mathbf{r}}{dt^2} = \frac{\mathbf{r}(t+\Delta t) - 2\mathbf{r} + \mathbf{r}(t-\Delta t)}{\Delta t} \tag{5.16}$$

$$\frac{d\mathbf{r}}{dt} = \frac{\mathbf{r}(t+\Delta t) - \mathbf{r}(t-\Delta t)}{2\Delta t}. \tag{5.17}$$

Substituting this into Equation (5.14) we have

$$\mathbf{r}(t+\Delta t) = \mathbf{r}(t) + (\mathbf{r}(t) - \mathbf{r}(t-\Delta t))\frac{1-\frac{1}{2}\gamma\Delta t}{1+\frac{1}{2}\gamma\Delta t} + \left(\frac{\Delta t^2}{m}\right)\frac{\mathbf{F}(t)+\mathbf{A}(t)}{1+\frac{1}{2}\gamma\Delta t}. \tag{5.18}$$

As $\gamma \to 0$, $\mathbf{A}(t) \to 0$ and Equation (5.18) reduces to the Verlet scheme, Equation (5.3).

To obtain the leapfrog Verlet equivalent expression we simply need to substitute in Equations 5.8 and 5.9 into Equation (5.18):

$$\mathbf{v}(t+\Delta t/2) = \mathbf{v}(t-\Delta t/2)\frac{1-\frac{1}{2}\gamma\Delta t}{1+\frac{1}{2}\gamma\Delta t} + \frac{\Delta t}{m}\frac{\mathbf{F}(t)+\mathbf{A}(t)}{1+\frac{1}{2}\gamma\Delta t}. \tag{5.19}$$

Once again, as $\gamma \to 0$ the equation reduces to the leapfrog Verlet algorithm (Equation (5.10)).

5.2 Ensembles

When simulating a biological system using MD it is important to use a biologically realistic environment. This means not only boundary conditions but also effects due to pressure and temperature. Basic MD maintains a microcanonical ensemble (NVE) while typically biological conditions are better described by an isobaric/isothermal ensemble (NPT). Several methods for maintaining constant temperature and pressure have been developed. Here we review a number of popular methods for temperature and pressure control available in common modelling packages and used in this thesis.

5.2.1 Berendsen Temperature Coupling

A number of methods have been developed to maintain constant temperature in MD simulations. The earliest methods were typically some form of *ad hoc* velocity rescaling [101, 106, 109] in which the velocitys of the atoms were simply rescaled to maintain a rigorously constant kinetic energy and, therefor, temperature. While straightforward and simple, this does not correspond to a constant temperature ensemble but a constant kinetic energy ensemble. Another method frequently used in this thesis and elsewhere is LD (see Section 5.1.4). This rigorously provides a constant temperature ensemble but perturbs the dynamics of the system through its stochastic nature and friction term. Two deterministic methods have been developed to maintain constant temperature: the Nosé-Hoover [112, 112–114] and Berendsen thermostats [115]. The Nosé-Hoover thermostat offers a true constant temperature ensemble without disturbing the dynamics of the simulation. However, it can create hotspots in the simulation, particularly problematic for implicit solvent simulations, and produces oscillations in the temperature when the reference temperature is changed. It was also pointed out by Hoover that the method is not ergodic. Fortunately, this can be remedied by connnecting multiple thermostats in a chain.

The Berendsen thermostat is a velocity rescaling method of weak coupling to an external bath which uses 'the principle of least local perturbation consistent with the required global coupling' [115]. That is, the velocities are changed minimally and gently to achieve the desired coupling. Physically, the method approximates the perturbations that would be observed in an ideal nonequilibrium experiment. Method is corresponds to an ensemble between the microcanonical and canonical ensembles. Furthermore, it has been shown that the fluctuations can be interpreted for either the canonical or microcanonical ensemble [116]. This method has the advantage that it has continuous dynamics and that the coupling strength can be adjusted. As the coupling parameter is reduced to zero, the simulation approaches the microcanonical ensemble. Furthermore, the effect of the coupling can be evaluated and controlled.

We begin by considering a system coupled to a heat bath of temperature T_0. This can be done by using the Langevin equation which we have already seen (Equation (5.14)) in 1D with no loss of generality

$$m_i\dot{v}_i = F_i(t) - \gamma_i m_i v_i + A(t) \tag{5.20}$$

where

$$\langle A_i(t)A_j(t')\rangle = 2m_i\gamma_i k_B T_0 \delta(t-t')\delta_{ij}. \tag{5.21}$$

The third term on the R.H.S. of Equation (5.20) corresponds to random noise while the second term is the coupling to the external heat bath. The strength of the coupling is determined by γ_i. We choose $\gamma_i = \gamma$ as, in practice, it does not affect the results. In principle one can couple different parts of the system to the heat bath with different strengths.

For a system under the influence of stochastic coupling the change in the kinetic energy w.r.t. time can be expressed via a simple finite difference

$$\begin{aligned}\frac{dE_k}{dt} &= \lim_{\Delta t\to 0}\frac{\sum_{i=1}^{3N} 1/2m_i v_i^2(t+\Delta t) - \sum_{i=1}^{3N} 1/2m_i v_i^2(t)}{\Delta t} \\ &= \sum_{i=1}^{3N}\lim_{\Delta t\to 0}\frac{m_i\Delta v_i^2}{2\Delta t} + \sum_{i=1}^{3N} m_i v_i(t)\dot{v}_i \\ &= 3N\gamma k_B T_0 + \sum_{i=1}^{3N} m_i v_i(t)\dot{v}_i. \end{aligned}\tag{5.22}$$

In order to perturb the local system as little as possible we wish to remove the random noise of the third term of Equation (5.20) while maintaining the global coupling of the second term. Our modified equation of motion is then

$$m_i\dot{v}_i = F_i + m_i\gamma\left(\frac{T_0}{T}-1\right)v_i \tag{5.23}$$

where we have removed the stochastic term. Through calculating the time derivative of the kinetic energy using Equations (5.22) and (5.23),

$$\begin{aligned}\frac{dE_k}{dt} &= \sum_{i=1}^{3N}\lim_{\Delta t\to 0}\frac{m_i\Delta v_i^2}{2\Delta t} + \sum_{i=1}^{3N} m_i v_i(t)\dot{v}_i. \\ &= \sum_{i=1}^{3N} F_i v_i + 3Nk_B\gamma\left(T_0 - T\right). \end{aligned}\tag{5.24}$$

The first term of in the R.H.S. of this result is the negative change in potential energy, i.e. the change in kinetic energy without a heat bath. The second term is the change in energy due to the global coupling to the heat bath.

Equation (5.23) represents a smooth scaling of the velocities for each time step in a finite difference scheme. To first order this scaling from v to λv is

$$\lambda = 1 + \gamma\Delta t\left(\frac{T_0}{T}-1\right). \tag{5.25}$$

The temperature change per time step can be made exactly $(T_0 - T)2\gamma\Delta t$. This gives

$$\lambda = \left[1 + 2\gamma\Delta t\left(\frac{T_0}{T}-1\right)\right]^{1/2}. \tag{5.26}$$

There are some interesting properties in the system to note. The first is that $\sum m_i(\Delta v_i)^2$ is minimized while $\sum\Delta(1/2mv^2)$ is constrained. This means that we have a least squares

local disturbance that satisfies a global constraint. Furthermore, the Maxwellian shape of the velocity distribution is conserved.

As previously stated, this temperature coupling does not correspond to any known ensemble. This generally means that no useful equations can be given that will allow the use of measured fluctuations in the system. For nonequilibrium MD this is not a problem since such fluctuations are, typically, not meaningful. For equilibrium MD the coupling constant can be chosen such that the constraint placed on the system is negligible and merely prevents the temperature from ramping up or down. In such a case the appropriate constant energy ensemble may be used for calculations.

When using Berendsen temperature coupling, one must choose both the temperature of the heat bath and the value of the coupling constant. The former should be straightforward while the later depends on the type of simulation being done. A typical values used are $\gamma = 5$ ps^{-1} 0.1 ps^{-1}.

5.2.2 Pressure Coupling

The same approaches used for temperature coupling can also be used to perform pressure coupling. While the Berendsen weak coupling approach has been used for pressure, it generally suffers the same problems as the Berendsen thermostat [117]. A Langevin method has been produced [117] and has been implemented in NAMD. Here we describe a version of a Nosé-Hoover chain approach for both constant pressure and temperature by Martyna [118, 119].

This method makes use of the extended ensemble approach, in which additional dynamical variables are introduced. The equations of motion for a particle, i, proposed in this case are

$$\frac{d\mathbf{r}_i}{dt} = \frac{\mathbf{p}_i}{m_i} + \frac{p_\epsilon}{W}\mathbf{r}_i \tag{5.27}$$

$$\frac{d\mathbf{r}_i}{dt} = \mathbf{F}_i - \left(\frac{p_\epsilon}{W} + \frac{p_\varepsilon}{Q}\right)\mathbf{r}_i \tag{5.28}$$

$$\frac{dV}{dt} = \frac{DVp_\epsilon}{W} \tag{5.29}$$

$$\frac{dp_\epsilon}{dt} = DV\left(P_{\text{int}} - P_{\text{ext}}\right) - \frac{p_\epsilon p_\varepsilon}{Q} \tag{5.30}$$

$$\frac{d\varepsilon}{dt} = \frac{p_\varepsilon}{Q} \tag{5.31}$$

$$\frac{dp_\varepsilon}{dt} = \sum_{i=1}^{N}\frac{\mathbf{p}_i^2}{m_i} + \frac{p_\epsilon^2}{W} - (N_f + 1)kT \tag{5.32}$$

where we introduce the dimension, D, number of degrees of freedom, N_f, volume, V, barostat momentum p_ϵ, thermostat position, ε, and momentum p_ε. P_{ext} is the external/applied pressures while

$$P_{\text{int}} = \frac{1}{DV}\left[\sum_{i=1}^{N}\left(\frac{\mathbf{p}_i^2}{m_i} + \mathbf{r}_i \cdot \mathbf{F}_i\right) - DV\frac{\partial U(\mathbf{r}, V)}{\partial V}\right] \tag{5.33}$$

and U is the potential energy.

These equations should be coupled in a chain (see [118] and [119] for details). This yields a conserved quantity of

$$H' = \sum_{i=1}^{N}\frac{\mathbf{p}_i^2}{2m_i} + \frac{p_\epsilon^2}{2W} + \frac{p_\varepsilon^2}{2Q} + U(\mathbf{r}, V)kT\varepsilon + (N_f + 1) + P_{\text{ext}}V. \tag{5.34}$$

That is,

$$\frac{dH'}{dt} = 0. \tag{5.35}$$

5.3 Boundary Conditions

The objective of MD is to simulate a small number of particles ($\mathcal{O}(10^2) - \mathcal{O}(10^6)$) to understand the properties of a large number of particles ($\mathcal{O}(6.022 \times 10^{23})$. In moving from the microscopic to the macroscopic, boundary conditions become all important. There are three general types of boundary conditions: free boundary conditions (*in vacuo*), periodic boundary conditions (PBC) and various types of extended wall region boundary conditions. We will focus primarily on *in vacuo* and PBC methods.

5.3.1 Free Boundary Conditions

Free Boundary Conditions or *in vacuo* simulations correspond to the zero pressure gas phase of the system. Typically, the protein or DNA strand is solvated in a sphere of water. No constraints are put on the system and water is allowed to evaporate [106, 109].

The main advantage of this system is the computational simplicity. A minimum number of water molecules are used to hydrate the solute (globular proteins in particular) and spherical cutoffs may be used for non-bond interactions. This also means that a surface molecule will 'see' fewer atoms than a buried atom and this further reduces computation time.

A second advantage is that the system can be used to study polarization effects that are difficult to observe with other boundary potentials. That is, the water and ions react only to the protein and not to images or boundary interactions. Thus the polarization of the water is solely due to local effects

There are some serious drawbacks to this method. Surface effects play a much larger role than in a macroscopic system as the number of atoms on the surface is proportional to $N^{1/3}$. To illustrate the effect consider a simple cube of 1000 atoms, 49% of these are found on the surface. For 1,000,000 atoms the percentage of surface atoms drops to just 6%. For one mole of atoms this proportion drops to 9×10^{-6}%. Thus, surface tension plays an important role in the simulation and can seriously distort the shape of the water drop.

Another draw back is that the system does not, effectively, contain a constant number of particles. As particles evaporate and move beyond the spherical cutoff of their neighbour atoms they are no longer part of the system. This can lead to artifacts when trying to analyze various properties of the water.

5.3.2 Periodic Boundary Conditions

A solution to many of the problems of *in vacuo* simulation is the use of periodic boundary conditions (PBC). Here the system is solvated in an appropriately shaped box that represents the unit cell of an infinite crystal lattice [106, 109].

Generally, the sum of the infinite lattice interactions is not done (electrostatic Ewald summation is an exception, see Section 5.5.2). Rather the minimum image convention is used. Particle i only interacts with the nearest image of particle j and never interacts with its own images. The nearest image may be the real particle or an image. For this reason, when using spherical cutoff schemes, r_{off} must be less than half the smallest box side length of the system. This helps to prevent spurious correlations that periodicity may impose on the system compared to that of a truly bulk system.

Even with the minimum image convention box size and shape still must be chosen appropriately to minimize artifacts due to periodicity. This is particularly true for solid phase simulations as the box size and shape should fit the size and shape of the crystal is or is expected to be formed. Otherwise the crystal will be deformed or, more likely, a glass will be formed.

Another consideration when choosing box size and shape for proteins is reducing the amount of water that needs to be simulated. For a globular protein, a cubic or rectangular box typically has unnecessary water in the 'corners' of the simulation. By using a truncated octahedron unit cell these corners can be eliminated.

A final consideration is for the box size when dealing with a protein. For a cubic box the side length should be typically the maximum width of the protein plus twice the cutoff distance. This ensures that protein is not interacting with its images and the water is only directly interacting with one image of the protein. If Ewald summation is used, the full electrostatic interactions are calculated and the protein will interact with all of its images. The same convention can still be used though cutoffs (now only for van der Waals forces) are typically reduced.

5.4 Force Fields

At the heart of any MD simulation is the potential or force field used. No computational tricks or level of accuracy will produce a physically realistic simulation if a poor potential is used. A potential must be both simple, to make the computation feasible, and physically accurate. Modern force fields do a very good job but they are also the greatest limitation to physical accuracy in MD simulations.

All modern MD force fields are empirical. That is, they are made up of functions that approximate observed experimental equilibrium behaviour and parameterized based on experimental and *ab initio* simulation data. This results in simple potentials (largely made up of harmonic potentials) that reproduce most experimental data under equilibrium conditions.

The force fields we use are the CHARMM27 [101, 120] and Amber99,03 [121, 122] force fields. Both are similar in their approach with CHARMM27 having an extra Urey-Bradley corrective term. As with most empirical force fields, the potential energy is divided into bond and non-bond terms:

$$U = U_{\text{bond}} + U_{\text{non-bond}}. \tag{5.36}$$

U_{bond} describes chemical bonds between atoms and is further made up of bond distance, bond angle, Urey-Bradley 1,3-distance, dihedral angle (torsion) and improper torsion terms. $U_{\text{non-bond}}$ is made up of long range electrostatic and van der Waals terms.

$$U = U_{\text{bond}} + U_{\text{angle}} + U_{\text{UB}} + U_{\text{dihedral}} + U_{\text{improper}} + U_{\text{vdW}} + U_{\text{elec}}. \tag{5.37}$$

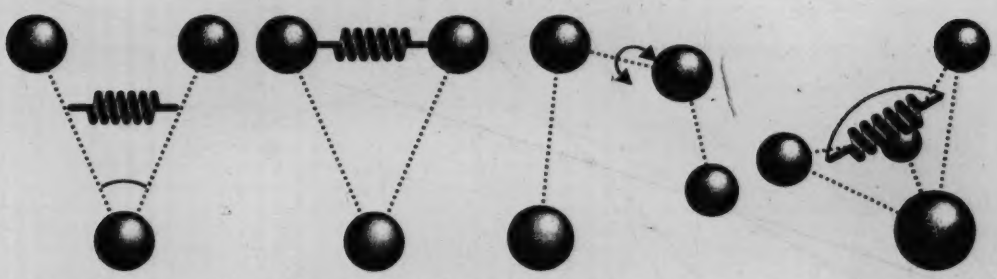
It should also be noted that the potentials are pair-wise and neglect N-body interactions.

5.4.1 Bond Distance

The distance between two chemically bound atoms is modeled as a simple harmonic potential (see Figure 5.1(e)).

$$U_{ij}^{\text{bond}}(b_{ij}) = K_{ij}^{b}(b_{ij} - b_{ij}^{0})^2 \tag{5.38}$$

b is the distance between the atoms, b^0 is the equilibrium distance and K^b is a empirically determined constant.

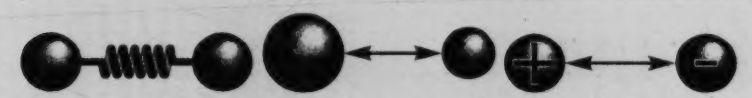


(a) Bond Angle. The angle between 1-3 atom pairs.

(b) Urey-Bradley. The distance between 1-3 atom pairs.

(c) Dihedral Torsion. The rotation between 1-4 atom bonds.

(d) Improper Torsion. The rotation between 2 planes of 3 atoms.



(e) Bond Distance. The distance between 1-2 atom pairs.

(f) van der Waals. The steric interactions between atoms separated by 3 or more bonds.

(g) Electrostatic. The electrostatic interactions between atoms separated by 3 or more bonds.

Figure 5.1: Graphical representation of potential energy terms. Bond distance (e), angle (a), Urey-Bradley (b) and improper dihedral (d) terms are approximated as harmonic potentials while dihedral torsions (c) are represented by a periodic potential. van der Waals (f) interactions are represented by the Lennard-Jones potential and electrostatics (g) are modeled with the coulomb potential.

5.4.2 Bond Angle

If three atoms are joined by bonds they form an angle (see Figure 5.1(a)). This angle is also modeled by a harmonic potential.

$$U_{ij}^{\text{angle}}(\theta_{ij}) = K_{ij}^{\theta}(\theta_{ij} - \theta_{ij}^{0})^2 \tag{5.39}$$

where θ is the angle of the bonds, θ^0 is the equilibrium angle and K^θ is an empirically determined constant.

5.4.3 Urey-Bradley

Urey-Bradley is primarily a corrective term applied when the existing parameters do not satisfactorily reproduce available vibrational spectra [120]. This takes the of an extra "bond" distance constraint between 1-3 atom pairs (atoms separated by two bonds, see Figure 5.1(b)).

$$U_{ij}^{\text{UB}}(S_{ij}) = K_{ij}^{S}(S_{ij} - S_{ij}^{0})^2 \tag{5.40}$$

S is the distance between the atoms, S^0 is the equilibrium distance and K^S is an empirically fit constant. This potential is not applied to all 1-3 pairs.

5.4.4 Dihedral Angle

Twisting or rotating a bond is represented by the dihedral angle (torsion) potential. This models steric barriers between 1-4 atom pairs (separated by 3 bonds) as a rotation about the central bond 5.1(c). This bond is assumed to be periodic and is usually represented by a cosine

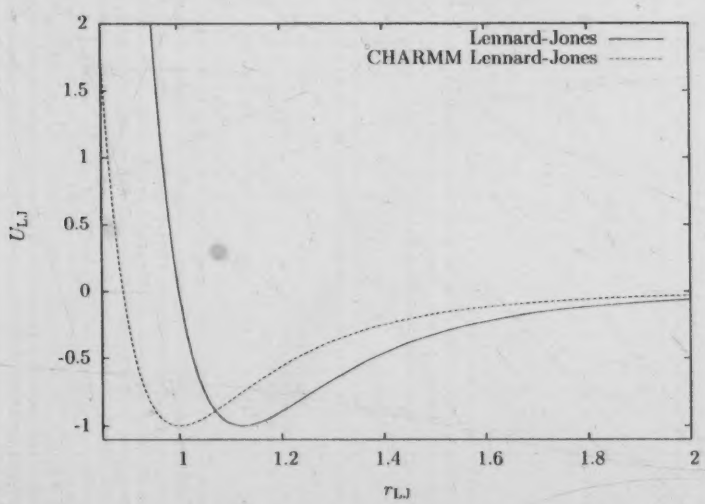


Figure 5.2: Comparison of the standard and CHARMM versions of the Lennard-Jones potential. $\epsilon = \sigma = \tilde{\sigma} = 1$.

$$U_{ij}^{\text{diheral}}(\phi_{ij}) = K_{ij}^{\phi}(1 + \cos(n_{ij}\phi_{ij} - \phi_{ij}^{0})),\ n = 1, 2, 3, 4, 6 \tag{5.41}$$

where ϕ is the angle, n is the coefficient of symmetry, ϕ^0 is an offset angle and K_{ij}^{ϕ} is an empirical constant.

5.4.5 Improper Torsion

Improper torsion is another corrective term. It is applied to groups of four atoms, $A-B-C-D$, where the angle between planes ABC and BCD (see Figure 5.1(d)) has the potential

$$U_{ij}^{\text{improper}}(\omega_{ij}) = K_{ij}^{\omega}(\omega_{ij} - \omega_{ij}^{0})^2 \tag{5.42}$$

where ω is the angle between the planes, ω^0 is the equilibrium angle and K^{ω} is a constant.

5.4.6 van der Waals

van der Waals forces represent the steric and induced dipolar interactions between atoms (see Figure 5.1(f)). This is typically represented by a Lennard-Jones 6-12 potential:

$$U_{ij}^{\text{LJ}}(r_{ij}) = 4\epsilon\left[\left(\frac{\sigma}{r_{ij}}\right)^{12} - \left(\frac{\sigma}{r_{ij}}\right)^{6}\right]. \tag{5.43}$$

This has a well depth of ϵ at $r_{ij} = 2^{1/6}\sigma$. σ is often called the Lennard-Jones radius as $U_{\text{LJ}}(\sigma) = 0$ (Figure 5.2). The r^{-6} term represents the dipole induced by the pair atom while the r^{12} term approximates the repulsion of the electrons due to the Pauli Exclusion Principle.

Both CHARMM and Amber use a slightly different form of the Lennard-Jones potential:

$$U_{ij}^{\text{LJ}}(r_{ij}) = \epsilon\left[\left(\frac{\tilde{\sigma}}{r_{ij}}\right)^{12} - 2\left(\frac{\tilde{\sigma}}{r_{ij}}\right)^{6}\right]. \tag{5.44}$$

ϵ is still the well depth but this now occurs at $r_{ij} = \tilde{\sigma}$. This also has the effect of changing the x-intercept to $r_{ij} = 2^{-1/6}\tilde{\sigma}$ (Figure 5.2). The main reason for this change is that the force is slightly more efficient to calculate.

$$F_{ij}^{\mathrm{LJ}}(r_{ij}) = -\frac{dU_{ij}^{\mathrm{LJ}}(r_{ij})}{dr_{ij}} = 24\epsilon\left(2\frac{\sigma^{12}}{r_{ij}^{13}} - \frac{\sigma^{6}}{r_{ij}^{7}}\right) \tag{5.45}$$

$$F_{ij}^{\text{CHARMM LJ}}(r_{ij}) = -\frac{dU_{ij}^{\text{CHARMM LJ}}(r_{ij})}{dr_{ij}} = 12\epsilon\left(\frac{\tilde{\sigma}^{12}}{r_{ij}^{13}} - \frac{\tilde{\sigma}^{6}}{r_{ij}^{7}}\right) \tag{5.46}$$

An important aspect of the Lennard-Jones potential in biomolecular simulations is the interaction between atoms of different types. The formula is well defined for systems with one type of atom, for example argon. However, when atoms such a oxygen and hydrogen interact the differences in well depth and radius must be accounted for. Some force fields have used the method of defining an interaction radius and well depth for each possible combination of atoms. This is done only in special cases for CHARMM. The standard method employed, the Lorentz-Berthelot mixing rules, to take the geometric average of the well depth and the arithmetic mean of the radius [120]. That is,

$$\epsilon_{ij} = \sqrt{\epsilon_i \epsilon_j} \tag{5.47}$$

$$\sigma_{ij} = \frac{\sigma_i + \sigma_j}{2} \tag{5.48}$$

It should also be noted CHARMM does not do Lennard-Jones calculations for 1-2 and 1-3 atom pairs (atoms separated by one and two bonds respectively). 1-4 interactions may have different parameters specified from the ones generally used in non-bond interactions[120]. In Amber the 1-4 interactions are scaled by a factor of 0.5.

5.4.7 Electrostatics

Electrostatic interactions are arguably the most important interactions in biophysics. They dominate over van der Waals forces for long range inter-molecular interactions and they play significant role in non-chemical binding.

Within a molecule charge is not localized to specific atoms. However, the quantum mechanical calculation necessary to determine the distribution of charge is not feasible for each time step of the simulation. Therefore, each atom in a structure is assigned a fixed partial charge, typically a fraction of an electron even if the structure is neutral, determined through fitting to electrostatic potentials produced from *ab initio* calculations. This approximates the electric field produced by the structure and allows a degree of polarization through the movement of atoms. These charges are fixed at the time of parameterization of the force field.

As a result, the electrostatic interactions are between point charges (see Figure 5.1(g)) have the form

$$U_{ij}^{\mathrm{elec}}(r_{ij}) = w_{ij}\frac{q_i q_j}{\epsilon_{\mathrm{el}} r_{ij}} \tag{5.49}$$

Here ϵ_{el} is the effective dielectric constant and should typically be set to 1. Charges, q, are expressed in elementary charges divided by 18.2223, $e/18.2223$. w_{ij} is a weighting factor being 0 for 1-2 and 1-3 interactions. $w_{ij} = 1$ for 1-4 interactions in CHARMM but is scaled by 0.5 for Amber.

5.5 Calculation of Non-Bonded Interactions

Non-bond forces represent the bulk of the computation time for energy and force calculations. Two approaches are generally taken. Since the magnitude of the potential decreases radially the standard approach has been to use a spherical cutoff where non-bond contributions are ignored past a given separation. While only having been implemented in 1980s for biophysical calculations, the Ewald summation method predates computer simulations and can calculate the exact electrostatic potential for a periodic system. Both methods have their advantages and disadvantages and care must be taken to ensure that the simulations remain as physically realistic as possible.

5.5.1 Spherical Cutoffs

While Ewald summation, particularly particle-mesh Ewald (PME), is now standard for simulations with periodic boundary condition, non-periodic simulations (e.g. implicit solvation, infinite dilution) cannot take advantage of it. Thus, the spherical cutoff approximation remains important for coulomb interactions in non-periodic simulations but also for LJ interactions for all simulations.

For both the Lennard-Jones and electrostatic potentials the spherical cutoff method is based on the idea that the non-bond potentials converge to 0 as $r \to \infty$.

5.5.1.1 Lennard-Jones Cutoff Schemes

Since Lennard-Jones interactions are dominated by the r^{-6} term at long ranges, this potential converges quite quickly and is well suited to a cutoff scheme. However, simple truncation at a given distance causes a discontinuity in the potential, creating large forces at the cutoff distance. Furthermore, the discontinuity in the force means that total energy is not conserved during the calculation[123]. We can see this from the work-energy theorem where we have $W_{\text{net}} = \Delta K$. Taking the force on the particle at the cutoff to be $\mathbf{F}(r_{\text{off}}) = \mathbf{F}_{\text{off}}$ when moving from inside the cutoff to outside we have

$$\begin{aligned} \Delta K &= \mathbf{F}_{\text{off}} \cdot \Delta \mathbf{r} \\ &= \mathbf{F}_{\text{off}} \cdot \mathbf{v} \Delta t. \end{aligned}$$

However, when the reverse is true the particle experiences no force and $\Delta K = 0$. It can be seen from this that as $\Delta t \to 0$, $\Delta K \to 0$. Thus, a small enough time step can provide adequate energy conservation, depending on the cutoff radius.

The effect is not as pronounced in Metropolis simulations as the force is not directly used. However, there is an impulsive contribution to the pressure that must be accounted for[106]. As such, two methods are typically used to remove the discontinuity [1]: the shift potential and the switch potential.

Potential Shift The shift potential modifies the pure Lennard-Jones potential by adding to the potential a term $U_{\text{LJ}}(r_{\text{off}}) = 0$. Given the definition, there are a number of different ways of achieving this.

[1] Of the four packages discussed here, only Amber does not support 'smooth' potentials or forces for LJ and coulomb interactions [98].

A common method is to add a constant term to the potential [106]. This gives a potential of the form

$$U_{\text{sh}}^{\text{LJ}}(r) = \begin{cases} 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - 2\left(\frac{\sigma}{r}\right)^{6} - \left(\frac{\sigma}{r_{\text{off}}}\right)^{12} + 2\left(\frac{\sigma}{r_{\text{off}}}\right)^{6}\right] & r < r_{\text{off}} \\ 0 & r > r_{\text{off}}. \end{cases} \tag{5.50}$$

The shift has the effect of raising the potential such that at r_{off} the untruncated potential is zero (Figure 5.3(b)). Since we are free to change the potential energy by an arbitrary constant this has no effect on the dynamics of the system for particles within the cutoff. Thus, this form of the potential also has minimal effect on the force other than the missing tail force as can be seen the Figure 5.3(c). Unfortunately, the discontinuity of the force means that this potential does not conserve energy.

An alternative method, used by CHARMM, is to add to the potential the function $S(r) = Cr^6 + D$[123]. C and D are chosen for each atom pair such that both the potential and force are zero at $r = r_{\text{off}}$ and ensures that the conservative nature of the potential is not lost due to a discontinuity. Thus, we have

$$U_{\text{CHARMM sh}}^{\text{LJ}} = \begin{cases} \epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - 2\left(\frac{\sigma}{r}\right)^{6} + Cr^6 + D\right] & r < r_{\text{off}} \\ 0 & r > r_{\text{off}} \end{cases} \tag{5.51}$$

$$F_{\text{CHARMM sh}}^{\text{lj}} = \begin{cases} 6\epsilon\left[2\left(\frac{\sigma^{12}}{r^{13}}\right) - 2\left(\frac{\sigma^{6}}{r^{7}}\right) + Cr^5\right] & r < r_{\text{off}} \\ 0 & r > r_{\text{off}} \end{cases} \tag{5.52}$$

where $C = 2\left(\frac{\sigma^{12}}{r_{\text{off}}^{18}} - \frac{\sigma^{6}}{r_{\text{off}}^{12}}\right)$ and $D = \frac{4\sigma^6}{r_{\text{off}}^6} - \frac{3\sigma^{12}}{r_{\text{off}}^{12}}$.

The main drawback of this method is that the potential for close interactions is modified. Most seriously, this changes the minimum of the system. However, for a large enough cutoff this is minimal.

Potenial Switch The preferred method for handling the Lennard-Jones interactions is with a switching function. Here, an additional distance less than the truncation cutoff, r_{on}, is specified. Between r_{on} and r_{off} a switching function is applied that smoothly reduces the potential to zero (Figure 5.3(b))[101].

$$U_{\text{sw}}^{\text{LJ}} = \begin{cases} 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - 2\left(\frac{\sigma}{r}\right)^{6}\right] & r < r_{\text{on}} \\ 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - 2\left(\frac{\sigma}{r}\right)^{6}\right]\left[\frac{(r_{\text{off}}^2 - r^2)^2(r_{\text{on}}^2 + 2r^2 - 3r_{\text{off}}^2)}{(r_{\text{off}}^2 - r_{\text{on}}^2)^3}\right] & r_{\text{on}} < r < r_{\text{off}} \\ 0 & r > r_{\text{off}}. \end{cases} \tag{5.53}$$

The advantage of this form is that it provides the exact potential and force until the distance r_{on}. Switching also has the benefit that the force is continuous. The trade-off for this is that when the potential is switched off a repulsive force results (Figure 5.3(c)). The magnitude of this force should be small since r_{off} should be large enough to minimize other truncation effects. It can be further reduced by making the switching region larger. Switching is the default method used in CHARMM [124] and NAMD [125].

In choosing r_{off} for out simulation we wish to minimize any truncation effects. A typical cutoff is $r_{\text{off}} = 2.5\sigma$ which corresponds to approximately 1/60th of the well depth for a standard Lennard-Jones potential (about 1/120th of the CHARMM potential well depth)[106]. For an atom such as TIP3P oxygen with a Lennard-Jones radius of $r_{\text{LJ}} = 3.5364$Å this corresponds to $r_{\text{off}} = 8.841$Å. Since we use the same cutoff distances for the Lennard-Jones and electrostatic potentials we typically use $r_{\text{on}} = 11$Å and $r_{\text{off}} = 15$Å or $r_{\text{off}} \approx 4.24\sigma$ which gives $U_{\text{LJ}}(r_{\text{off}})/U_{\text{LJ}}(\sigma) \approx 1/3000$.

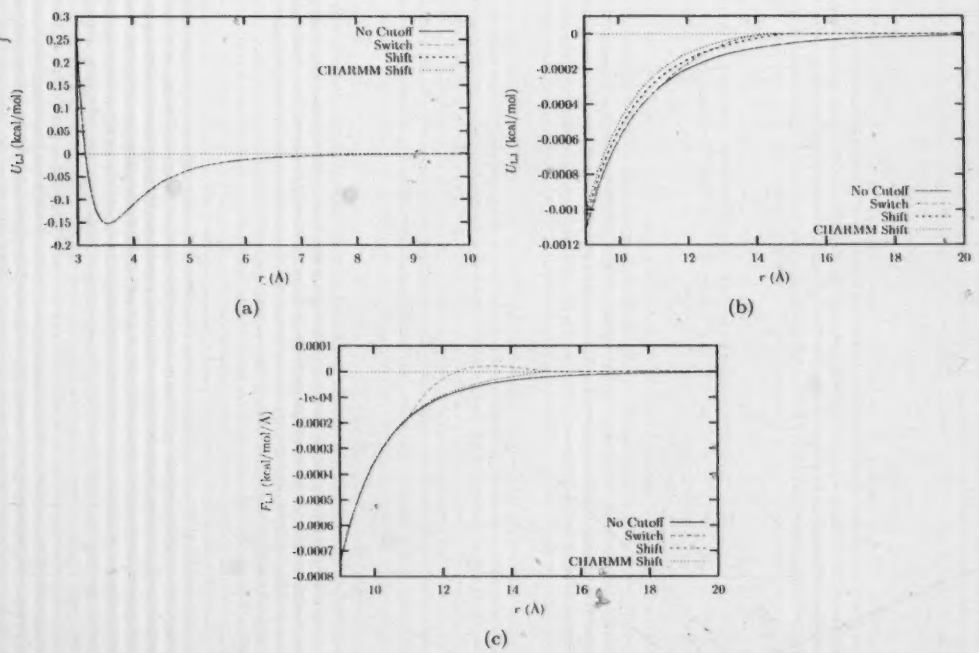


Figure 5.3: Lennard-Jones cutoff schemes applied to TIP3P water oxygen-oxygen interactions with $r_{on} = 11$Å and $r_{off} = 15$Å. Neither shifting nor switch the potential appears to have much effect on the potential (a), taking a closer look at the cutoff region reveals the differences between the three schemes (b). The resulting force in the cutoff region is shown in (c). We can see that the switch potential creates a repulsive while the shifted potential has a discontinuity in the force.

5.5.1.2 Electrostatic Interactions

Electrostatic interactions are particularly difficult to efficiently calculate as they converge slowly with respect to r. A number of schemes have been developed to deal with these forces in an efficient manner [119]. Radial cutoffs remain the simplest and, often, the fastest method.

Many studies have been done that discuss various forms of the spherical cutoff (see, for example, [123] and [126]). Due to the variety of methods and the previous work that has been done, we will briefly discuss only a few notable examples: truncation, shift, force shift, switch and force switch.

It should also be noted that when determining whether or not an atom lies within the cutoff region two methods are used: group based and atom based. In the atom based method the center of the atom is used to decide if the atom in inside the cutoff region. The group based method uses the center of mass of a predetermined group of atoms to decide whether or not an individual atom is inside or outside the cutoff. The groups are chosen to have an integer charge and typically consist of about five atoms. This approach is deeply embedded in the CHARMM force field, for example [120].

Simple Truncation As with Lennard-Jones interactions, simple truncation involves calculating the true electrostatic interactions until a separation of r_{off} is reached. The same issues with conservation of energy arise though they are typically more severe since the electrostatic potential drops off much more slowly than does the Lennard-Jones potential. In addition to this there is a dipole heating effect that is independent of step size associated with group truncation[123]. Typical solutions to this involve a strong coupling to a heat bath but this merely masks the unphysical effects.

Potential Switch Another cutoff scheme that is problematic is the potential switch method. On the surface this appears to be a very reasonable method and was used for a long period of time in various forms. As in the Lennard-Jones potential switch (Equation (5.53)) the exact potential is calculated until a separation of r_{on}. After this point a switching potential, of the form $S(r) = \left[\frac{(r_{\text{off}}-r)^2(r_{\text{off}}+2r-3r_{\text{off}})}{(r_{\text{off}}-r_{\text{on}})^3}\right]$, is invoked until a separation of r_{off}.

$$U_{\text{sw}}^{\text{elec}} = \begin{cases} \frac{q_i q_j}{\epsilon_{\text{el}} r} & r < r_{\text{on}} \\ \frac{q_i q_j}{\epsilon_{\text{el}} r}\left[\frac{(r_{\text{off}}^2-r^2)^2(r_{\text{off}}+2r-3r_{\text{on}})}{(r_{\text{off}}^2-r_{\text{on}}^2)^3}\right] & r_{\text{on}} < r < r_{\text{off}} \\ 0 & r > r_{\text{off}}. \end{cases} \tag{5.54}$$

$$F_{\text{sw}}^{\text{elec}} = \begin{cases} \frac{2q_i q_j}{\epsilon_{\text{el}} r^2} & r < r_{\text{on}} \\ \frac{q_i q_j}{\epsilon_{\text{el}} r}\frac{(r_{\text{off}}^2-r^2)^3}{(r_{\text{off}}^2-r_{\text{on}}^2)^3} \\ \left[\frac{(r_{\text{off}}+2r-3r_{\text{on}})}{r^2} + \frac{4(r_{\text{off}}+2r-3r_{\text{on}})}{(r_{\text{off}}^2-r^2)^2} - \frac{4}{(r_{\text{off}}^2-r^2)}\right] & r_{\text{on}} < r < r_{\text{off}} \\ 0 & r > r_{\text{off}}. \end{cases} \tag{5.55}$$

This switching function has the property that $S(r_{\text{on}}) = 1$, $S(r_{\text{off}}) = 1$, $dS/dr(r_{\text{on}}) = 0$ and $dS/dr(r_{\text{off}}) = 0$. This provides a continuous function for both potential energy and force (see Figure 5.4), meaning that the total energy is conserved. Though this function is specific to CHARMM the properties are generally true of all switching functions.

The potential switch for electrostatics suffers from the same weakness as does the potential switch for Lennard-Jones interactions. However, due to the relatively large forces and potentials encountered at the cutoff distance the resulting artificial forces are considerably greater as shown in Figure 5.4(b). The cutoff and switching region used in Figure 5.4 are considerably larger than what was recommended prior to the mid-1990s (e.g. $r_{\text{on}} = 7.5$Å and $r_{\text{off}} = 8.0$Å)

so the unphysical forces were typically even larger[123]. Furthermore, simulations of MbCO dynamics have shown that switching regions shorter than 4Å artificially reduce protein motion. Using the switch method with $r_{on} = 11$Å and $r_{off} = 12$Å Steinbach and Brooks observed a rms error in the Coulomb force of 31.35 kcal/mol/Å compared to 12.18 kcal/mol/Å when completely ignoring the electrostatics. This was reduced considerably to 4.43 kcal/mol/Å when r_{on} was changed to $r_{on} = 8$Å, though it was still relatively high compared to other methods.

Potential Shift An early alternative to the potential switch method and default method in NAMD [125] was the potential shift. In this method the electrostatic potential energy is multiplied by a factor $S(r) = (1-(r/r_{off})^2)^2$ such that both the potential and force are continuous functions of r.

$$U_{sh}^{elec} = \begin{cases} \frac{q_i q_j}{\epsilon_{el} r}(1-(\frac{r}{r_{off}})^2)^2 & r < r_{off} \\ 0 & r > r_{off}. \end{cases} \tag{5.56}$$

$$F_{sh}^{elec} = \begin{cases} \frac{q_i q_j}{\epsilon_{el} r}(1-(\frac{r}{r_{off}})^2)(1+3(\frac{r}{r_{off}})^2) & r < r_{off} \\ 0 & r > r_{off}. \end{cases} \tag{5.57}$$

This, too, has the property of conserving energy and does not create any unphysical forces. However, it overestimates the potential energy and underestimates the forces involved in the interaction. This skewing of the short range force is undesirable but has far less impact than the spurious force generated by the potential switch.

Force Shift Force shift is similar to potential shift. The form of the shifting term is such that the Coulomb force is offset by a constant amount (see Figure 5.4(b)).

$$U_{fsh}^{elec} = \begin{cases} \frac{q_i q_j}{\epsilon_{el} r}(1-(\frac{r}{r_{off}}))^2 & r < r_{off} \\ 0 & r > r_{off}. \end{cases} \tag{5.58}$$

$$F_{sh}^{elec} = \begin{cases} \frac{q_i q_j}{\epsilon_{el}}(\frac{1}{r^2}-\frac{1}{r_{off}^2}) & \\ 0 & r > r_{off}. \end{cases} \tag{5.59}$$

Unlike the potential shift, the force shift underestimates both the potential and the force of the interaction. However, in separate tests using water and protein made up of neutral groups the force shifted potential out performed the potential shift noticeably and more accurately determined the force[123].

Force Switch Force Switch uses the same switching function as the potential switch, $S(r) = \left[\frac{(r_{off}-r)^2(r_{off}+2r-3r_{off})}{(r_{off}-r_{on})^3}\right]$, except that it is applied to the force and not the potential[123]. While this provides the continuity of d^2U/dr^2 it means to obtain the potential we much integrate the force. That is,

$$U_{fsw}^{elec} = \begin{cases} U^{true}(r) + \Delta U & r < r_{on} \\ -\int_{r_{off}}^{r} S(r')F^{true}(r')dr' & r_{on} < r < r_{off} \\ 0 & r > r_{off}. \end{cases} \tag{5.60}$$

ΔU is determined by the need for $U_{fsw}^{elec}(r)$ to be continuous. Carrying out this integration

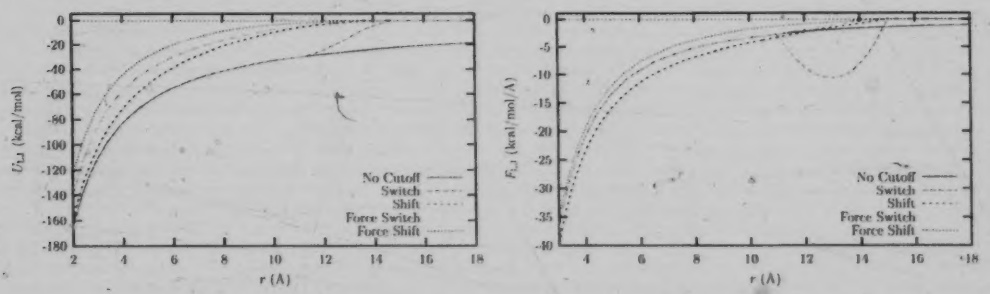


Figure 5.4: Electrostatic cutoff schemes applied to two point charges of $1e$ and $-1e$ with $r_{on} = 11$Å and $r_{off} = 15$Å. All of the cutoff schemes over estimate the potential energy for short range interactions except for the potential switching method(a). However, when looking at the forces both the potential switch and force switch methods work equally well for short range interaction(b). During the switching interval the potential switch creates a large spurious force.

we arrive at the expressions for potential and force

$$U_{\text{fsw}}^{\text{elec}} = \begin{cases} \frac{q_i q_j}{\epsilon_{el}} \left(\frac{1}{r} + 8\frac{(r_{on}r_{off})^2(r_{off}-r_{on}) - (r_{off}^5 - r_{on}^5)/5}{\gamma} \right) & r < r_{on} \\ \frac{q_i q_j}{\epsilon_{el}} (\quad A(r^{-1} - r_{off}^{-1}) + B(r_{off} - r) & \\ \qquad +C(r_{off}^3 - r^3) + D(r_{off}^5 - r^5)) & r_{on} < r < r_{off} \\ 0 & r > r_{off} \end{cases} \tag{5.61}$$

$$F_{\text{fsw}}^{\text{elec}} = \begin{cases} \frac{q_i q_j}{\epsilon_{el} r^2} & r < r_{on} \\ \frac{q_i q_j}{\epsilon_{el} r^2} \left[\frac{(r_{off}-r)^2(r_{off}+2r-3r_{off})}{(r_{off}-r_{on})^3} \right] & r_{on} < r < r_{off} \\ 0 & r > r_{off} \end{cases} \tag{5.62}$$

where $A = r_{off}^4(r_{off}^2 - 3r_{on}^2)/\gamma$, $B = 6(r_{on}r_{off})^2/\gamma$, $C = -(r_{on} + r_{off})/\gamma$, $D = 2/(5\gamma)$ and $\gamma = (r_{off}^2 - r_{on}^2)^3$.

Although this method conserves energy for atom-based switching, it does not for group-based switching. The same rotating dipole problem, as discussed for simple truncation, applies here as well. Furthermore, energy is not conserved within the switching region.

One final consideration is the length of the switching region. If it is too short (1Å) artificial minima are introduced in the switching region as seen in the potential switch method, though typically much shallower. This can be remedied by increasing the switching region to 4Å. It was found by Steinbach and Brooks that this produced no artificial minima in simulations conducted with N-methyl acetamide[123].

While shift, force shift and force switch methods all work very well for systems composed of neutral groups (e.g. water) when large numbers of charged groups are present the force switch method is preferred. Force switch is the default method used in CHARMM.

5.5.2 Ewald Summation

Ewald summation and Particle Mesh Ewald (PME) summation offer full electrostatics with minimal time penalties above that of spherical cutoffs. This of course depends on the size of

the system and the cutoffs. The two methods give equivalent results within numerical error and the grid interpolation error associated with PME. PME is generally faster than Ewald.

Though Ewald and PME, in particular, have become the standard for many types of simulations the method is not without its drawbacks. PBC are required for the method as it calculates the total electrostatic potential felt by a charge in an infinite crystal. The point of using PBC in most biologically oriented molecular simulations is to attempt to simulate a bulk disordered system. To achieve this, system sizes are chosen such that molecules cannot interact with their own image. However, using Ewald summation one eliminates electrostatic cutoffs. Thus, artifacts due to periodicity must always be a concern. Another concern is that of time. The Ewald algorithm is $\mathcal{O}(N^{3/2})$ and PME is $\mathcal{O}(N \log N)$ compared to $\mathcal{O}(N)$ for the spherical cutoff if properly implemented.

The original Ewald method was devised by Ewald in 1921 to study ionic crystals, long before the concept of molecular simulations was born. The method first found use in the late 1970s for simulation of ionic systems. In the mid 1980s the technique gained popularity for non-crystalline biological systems and gained further popularity in the mid 1990s with the development of PME [127, 128]. The original Ewald technique we will be discussing here has been widely used over the years and has, as a result, been discussed in many references[111, 129, 130].

The potential for an infinite periodic lattice of point charges is given by

$$U_i^{\text{elec}} = \frac{1}{2} \sum_{|\mathbf{n}|=0}^{\infty} {}' \sum_{j=1}^{N} \frac{q_i q_j}{\epsilon_{el} |\mathbf{r}_{ij} + \mathbf{n}|}, \tag{5.63}$$

where the prime indicates that the series does not include the interaction $i = j$ for $\mathbf{n} = 0$. The goal of Ewald summation is to take this slowly converging sum and force it to converge more quickly.

In order to do this two Gaussian charge distributions are added to the original distribution of point charges. This is then broken down into three distributions that are treated separately and illustrated in Figure 5.5. The ρ_{fourier} consists of replacing all the point charges, including the reference charge, with Gaussian charge distributions of width $\sqrt{2}/\kappa$.

$$\rho_{\text{Gauss}}(r) = q \left(\frac{\kappa^2}{\pi} \right)^{3/2} \exp(-(\kappa r)^2) \tag{5.64}$$

The second charge distribution, ρ_{real}, consists of the original point charge distribution with additional Gaussian charge distributions of opposite charge superimposed. The point charge q_i and its corresponding Gaussian are not included.

A third charge distribution, ρ_{self}, is a correction term for including a Gaussian for the reference charge in Fourier space.

We then have the following expression for the total potential

$$\phi = \phi_{\text{Fourier}} - \phi_{\text{self}} + \phi_{\text{real}}. \tag{5.65}$$

The potential energy calculated via Ewald summation is now

$$\begin{aligned} U_{\text{Ewald}} =& U_{\text{Fourier}} + U_{\text{real}} - U_{\text{self}} \\ =& \frac{1}{2\pi V} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{G \neq 0} \frac{q_i q_j}{G^2} \exp\left(-\pi^2 \mathbf{G}^2/\kappa^2\right) \exp\left(-2\pi i \mathbf{G} \cdot (\mathbf{r}_j - \mathbf{r}_i)\right) \\ &+ \sum_{|\mathbf{n}|=0}^{\infty} {}' \sum_{i,j=1}^{N} \frac{q_j q_j}{|\mathbf{r}_{ij} + \mathbf{n}|} \text{erfc}(\kappa |\mathbf{r}_{ij} + \mathbf{n}|) - \frac{\kappa}{\sqrt{\pi}} \sum_{i=1}^{N} q_i^2 \end{aligned} \tag{5.66}$$

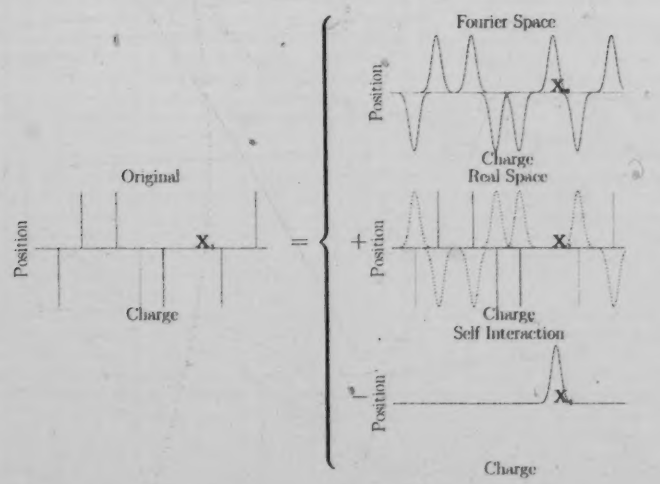


Figure 5.5: The original charge distribution, "Original", is reformulated as the sum of three other charge distributions. $\mathbf{X}_i$ indicates the reference point of the calculation. ρ_{Fourier} is illustrated in "Fourier Space". "Real Space" corresponds to ρ_{real}. The self interaction term, ρ_{self} is show in "Self Interaction".

where G is the reciprocal lattice vector and the prime, again, indicates that the series does not include the interaction $i = j$ for $\mathbf{n} = 0$.

In order for this to be a benefit to us it must converge quickly. This is controlled by the width of the Gaussian κ. The smaller κ is the faster the convergence in real space. However, the larger κ the faster the convergence in reciprocal space as fewer reciprocal space vectors must be used. Thus, a compromise is needed and this turns out to be $\kappa = 0.34$ and a maximum number of reciprocal vectors equal to $\kappa \times L_{(x,y,z)}$ where L is the box length[124].

Finally, since we cannot compute $\text{erfc}(x)$ exactly, it must be approximated. There are several methods of doing this. Among them are iterative techniques, Chebyshev polynomials, and linear or spline interpolation. Fifth order spline interpolation computed from a table of pre-calculated values provides the best balance of accuracy[124, 130].

PME improves upon the scaling of Ewald summation by starting with a clever choice of κ [127]. κ is chosen large enough that only atoms within a specified cutoff (typically 8 Å) contribute to U_{real} in Equation (5.66). U_{fourier} then includes only long-range contributions and the reciprocal space sum is approximated by a 3D, piecewise interpolated grid. The reciprocal space energy and forces are convolutions and are calculated in a straightforward manner with FFTs. The reciprocal sum is now the dominant term and scales as $N \ln(N)$ where N is the number of point charges.

5.5.3 Non-bonded List

As we have seen in Sections 5.5.1 and 5.5.2, applying cutoff distances to non-bond interactions is still an important aspect in computational efficiency. Not including the cost of evaluating whether or not an atom falls within a cutoff, MD simulations scale as $\mathcal{O}(N)$ rather than $\mathcal{O}(N^2)$ when a cutoff is used. To avoid unnecessarily evaluating distances for the cutoff, a cutoff, or

non-bonded, list is used. Since atoms move over the course of a simulation, the list needs to include atoms that are close to but outside the cutoff distance. Two basic methods have been created to produce such lists [119]: the Verlet list and the cell list.

The Verlet list is very simple in its approach. For each particle a list is made of all atoms within radius, $r_{\text{Verlet}} > r_{\text{cutoff}}$. Distances for all possible atom pairs are evaluated in constructing this list but the difference between r_{Verlet} and r_{cutoff} (1 - 2 Å) is chosen such that the list only needs to be constructed every ten to twenty time steps in most cases. Regardless, the method still scales as $\mathcal{O}(N^2)$.

$\mathcal{O}(N^2)$ scaling can be avoided by dividing the system into a grid with a linear spacing $\geq r_{\text{cutoff}}/2$. Each particle is assigned to a cell, an $\mathcal{O}(N)$ operation. When evaluating the distance, only cells within r_{cutoff} need be considered. More particles outside r_{cutoff} are considered for non-bonded interactions than in the Verlet list. However, since the overall scaling is far superior, cell list are preferred as soon as there are more than a few hundred particles.

Frenkel and Smit [119] detail hybrid cell-Verlet list methods that are superior to both methods.

5.6 Advanced Sampling Techniques

Increasing computing power has allowed the simulation of larger and more detailed systems. Nanosecond or microsecond time scales for systems of hundreds of thousands of atoms are now accessible. However, complete folding or sampling of even small protein fragments can be intractable due to the multiple potential minima that may trap the simulation. Similar problems have also plagued glassy systems. For this reason, several advanced sampling algorithms have been developed. The most successful of these are commonly known as generalized-ensemble algorithms[131–134]. These methods - multicanonical ensemble, simulated tempering and $1/k$ sampling - perform a random walk in potential energy, temperature and energy space respectively by using a non-Boltzmann probability weight. While extremely effective for biomolecular and other glassy systems, the probability weight is not known *a priori*. Rather the weighting is built up in an iterative process, typically a tedious and time consuming process.

Replica exchange, also known as parallel tempering and multiple Markov Chains, is another generalized-ensemble originally developed for Monte Carlo simulations [135–137] and is discussed in Section 5.6.1. It first use for MD was in dihedral space in 1997 [138]. This was adapted to Cartesian MD, replica exchange MD (REMD), in 1999 by Sugita and Okamoto [139] for the NVT ensemble and then to the NPT ensemble by Okabe *et al.* in 2001 [140]. As we shall see, the probability weight is known in advance for REMD, overcoming the main obstacle for generalized-ensemble algorithms. The method is also is naturally suited to a parallel computing architecture.

Another approach to sampling that has been the focus of consider interest is the calculation of potentials of mean force, or free energy profiles, along a specific reaction coordinate. Methods such as umbrella sampling [119], steered molecular dynamics with Jarzynski's equality [141, 142] and adaptive biasing force [143] attempt to uniformly sample the entire reaction coordinate, something that, with the exception of the trivial case of a flat profile, does not occur in standard MD or REMD. In Section 5.6.2 we discuss the adaptive biasing force MD in detail.

5.6.1 Replica Exchange in Temperature Space

We begin by considering a system of N particles, each with a position $\mathbf{q}_i$ and momentum $\mathbf{p}_i$ with a Hamiltonian

$$H = \sum_i^N K(\mathbf{p}_i) + E(\mathbf{q}_i). \tag{5.67}$$

For the NVT ensemble, each state $\mathbf{x} = \{\mathbf{q}_1 \dots \mathbf{q}_N, \mathbf{p}_1 \dots \mathbf{p}_N\}$ has a Boltzmann weight

$$\begin{aligned} W(\mathbf{x}, T) &= \exp\{-\beta H(\{\mathbf{q}_1 \dots \mathbf{q}_N, \mathbf{p}_1 \dots \mathbf{p}_N\})\} \\ &= e^{-\beta(E(\{\mathbf{q}_1 \dots \mathbf{q}_N\}) + \frac{3N}{2\beta}} \end{aligned} \tag{5.68}$$

where $\beta = 1/kT$ and T is the temperature.

We can now consider an ensemble of M non-interacting replicas of our system (only the values for x differ), each at a different temperature. The state of this ensemble is then $\mathbf{X} = \{\mathbf{x}_1 \dots \mathbf{x}_M, T_1 \dots T_M\}$ with a Boltzmann weight

$$\begin{aligned} W_{\text{Rep}}(\mathbf{X}) &= \Pi_i^M W(\mathbf{x}_i, T_i) \\ &\triangleq \exp\left\{-\sum_i^M \beta_i (E(\mathbf{x}_i) + \frac{3N}{2\beta_i}\right\}. \end{aligned} \tag{5.69}$$

At this point we can consider the probability of exchanging the temperatures (or another property) of two replicas, i.e.

$$\mathbf{X} = \{\dots, T_a, \dots, T_b, \dots\} \to \mathbf{X}' = \{\dots, T_b, \dots, T_a, \dots\}. \tag{5.70}$$

To maintain equilibrium and allow our system to converge, we impose a detailed balance condition

$$W_{\text{Rep}}(\mathbf{X}) w(\mathbf{X} \to \mathbf{X}') = W_{\text{Rep}}(\mathbf{X}') w(\mathbf{X}' \to \mathbf{X}). \tag{5.71}$$

This gives

$$\begin{aligned} \frac{w(\mathbf{X} \to \mathbf{X}')}{w(\mathbf{X}' \to \mathbf{X})} &= \frac{W_{\text{Rep}}(\mathbf{X}')}{W_{\text{Rep}}(\mathbf{X})} \\ &= \exp(-(E(\mathbf{q}_a) - E(\mathbf{q}_b))(\beta_b - \beta_a)) \\ &= e^{-\Delta}. \end{aligned} \tag{5.72}$$

This is be satisfied with the Metropolis criteria [144]

$$w(\mathbf{X} \to \mathbf{X}') = \begin{cases} 1 & \text{if } \Delta \leq 0 \\ e^{-\Delta} & \text{if } \Delta > 0 \end{cases} \tag{5.73}$$

The above treatment works for Monte Carlo or MD. However, for MD we need to additionally consider $\mathbf{p}_i$ after a successful exchange, as temperature and the average kinetic energy are related by

$$\langle K(\mathbf{p}) \rangle = \left\langle \sum_i^N \frac{\mathbf{p}_i^2}{2m_i} \right\rangle = \frac{3}{2} NkT. \tag{5.74}$$

Sugita and Okabe [139] suggest the transform

$$\begin{aligned} \mathbf{p}_a &\rightarrow \sqrt{\frac{T_b}{T_a}}\mathbf{p}_b \\ \mathbf{p}_a &\rightarrow \sqrt{\frac{T_b}{T_a}}\mathbf{p}_b \end{aligned} \quad , \tag{5.75}$$

to satisfy Equation (5.74).

For the NTP ensemble, Okabe *et al.* [140] show that the only change is that Δ becomes

$$\Delta = (E(\mathbf{q}_a) - E(\mathbf{q}_b))(\beta_b - \beta_a) + (\beta_b P_b - \beta_a P_a)(V_a - V_b) \tag{5.76}$$

where P and V are the pressure and volume.

While REMD is most commonly used and easily expressed as a random walk through temperature space, it is clear than any variable of the system can be used, as long as an exchange probability can be calculated.

5.6.1.1 Free Energy Calculations

Assuming that the simulation has properly simulated configurational space it is possible to calculate free energies along arbitrarily specified reaction coordinates using

$$A(\xi) = -\frac{1}{\beta}\ln \mathcal{P}_\xi + A_0, \tag{5.77}$$

where ξ is a point along the reaction coordinate, $\mathcal{P}_\xi$ is the probability of finding the system at that point and A_0 is an arbitrary constant. This method has been used to map the free energy folding landscapes of small peptides, such as Met-enkephalin [145]. However, as the landscape is not uniformly sampled, areas of high free energy also have high relative error. As a result, the heights of activation barriers are difficult to accurately predict.

5.6.1.2 Scaling

In practise, each replica is run on one or more CPUs and exchanges are attempted every few thousand time steps. As only temperatures are being exchanged, the communication frequency is low as is the bandwidth required. Furthermore, only neighbours in temperature space are considered for exchange as the probability of exchange drops exponentially with increasing temperature difference. This yields a naturally parallel algorithm that scales nearly linearly with the number of replicas.

While REMD scales linearly with the number of replicas, the number of replicas required for a temperature range generally scales as the square root of the number of degrees of freedom [146]. For all atom, explicit solvent simulations, where each simulation requires a separate CPU, the number of atoms that can be simulated is limited by the size of the parallel computer available. This is coupled with the fact that MD for each individual replica will scale, at best, linearly with the number of atoms, imposing another limit on system size.

With this limitation on system size, several attempts have been made to reduce the number of degrees of freedom. The simplest approach is to use implicit solvent. As well as removing the water degrees of freedom and the necessity of integrating their equations of motion, the lack of friction due to solvation also improves sampling. However, implicit solvent is often not an adequate replacement for explicit solvent and produces conformational artifacts and bias [77]. A hybrid approach has also been proposed [147]. In this case, an explicit solvent is used for

MD but is excluded from the potential energy calculations for exchange attempts. Instead, an implicit solvent is used in its place. This does not accelerate the MD part of the simulation but can drastically lower the number of replicas needed for the temperature range.

Under the weak coupling assumption, the degrees of freedom may be further reduced by focusing on only part of the solute [148]. Here, the target is typically a small part of the solute and is coupled to a different thermostat than the remainder of the system. For each replica, the main part of the solute has the same target temperature while the focus of REMD is has a different target temperature. The full potential energy is still used in calculating the exchange probability but the number of required replicas is reduced. As this has only been used with implicit solvent thus far, the number of replicas needed has been extremely low.

Multiplexed REMD (MREMD) does not attempt to reduce the degrees of freedom but targets the efficient use of large, widely distributed computing resources [149]. Traditional computing clusters are limited in size and can not handle REMD simulations of large systems. There exist millions of computers world wide that are often idle and that could potentially fill this need. However, REMD requires dedicated, homogeneous resources to work. One attempt of the Folding@Home project has been to create additional copies (multiplexes) of the REMD ensemble. Each replica is simulated as usual on a given computer. When a replica is ready for exchange, it is permitted to attempt an exchange with the appropriate temperature of any multiplex that also happens to be ready for exchange. In this way, heterogeneous collections of computers may be used and the impact of high latency is generally reduced as well. Further, if a compute node is lost, the replica can wait until another becomes available while the rest of the simulation continues.

5.6.2 Adaptive Biasing Force

A wide variety of methods have been developed to calculate free energies and differences in free energies [119, 150]. Most commonly, one is interested in a the free energy of the system at a state defined by a particular value of a reaction coordinate (RC), ξ:

$$A(\xi) = -\frac{1}{\beta} \ln \mathcal{P}_\xi + A_0, \tag{5.78}$$

where $A(\xi)$ is the free energy associated with the probability, $\mathcal{P}_\xi$, of finding the system at a given value of ξ. Often, it is more practical to calculate the difference between two states along an RC, defined as

$$\frac{dA(\xi)}{d\xi} = -\frac{1}{\beta \mathcal{P}_\xi} \frac{d\mathcal{P}_\xi}{d\xi}. \tag{5.79}$$

Suitably rearranged, this equation can be integrated to obtain a free energy profile along the reaction coordinate, sometimes called the potential of mean force (PMF).

While calculating the probabilities directly can be done, especially with advanced sampling techniques like replica-exchange [133], it is generally not efficient, especially if high activation barriers are involved. Low sampling of these regions leads to large errors. Other techniques, such as thermodynamics integration (TI) [119] and umbrella sampling (US) [119] have sought to overcome this by enforcing near uniform sampling over the entire range of interest of the RC. An alternative method for calculating Equation (5.79) is to treat ξ as a generalized coordinate, giving the relation

$$\frac{dA(\xi)}{d\xi} = \left\langle \frac{\partial U(\mathbf{x})}{\partial \xi} - \frac{1}{\beta} \frac{\partial \ln |J|}{\partial \xi} \right\rangle_\xi = -\langle F_\xi \rangle_\xi, \tag{5.80}$$

where $U(\mathbf{x})$ is the potential energy of the system, $\langle F_\xi \rangle_\xi$ is the ensemble-average for acting along the RC and $|J|$ is the determinant of the Jacobian for the inverse transformation from

generalized to Cartesian coordinates:

$$J = \begin{pmatrix} \partial x_1/\partial \xi & \partial x_1/\partial q_1 & \cdots & \partial x_1/\partial q_{3N-1} \\ \partial x_2/\partial \xi & \partial x_2/\partial q_1 & \cdots & \partial x_2/\partial q_{3N-1} \\ \vdots & \vdots & \ddots & \cdots \\ \partial x_{3N}/\partial \xi & \partial x_{3N}/\partial q_1 & \cdots & \partial x_{3N}/\partial q_{3N-1} \end{pmatrix}. \tag{5.81}$$

Equation (5.80) has been derived in the literature several times before for both the 1D case [150–153] and multidimensional case [154, 155]. In practice, the RC is divided into bins, Equation (5.80) is calculated for each bin and the result is numerically integrated to obtain the PMF.

In MD $\langle F_\xi \rangle_\xi$ is calculated as a time average with the instantaneous F_ξ given by

$$F_\xi = -\frac{\partial U(\mathbf{x})}{\partial \xi} + \frac{1}{\beta}\frac{\partial \ln |J|}{\partial \xi} \tag{5.82}$$

$$= -\sum_{k=1}^{3N} \frac{\partial U(\mathbf{x})}{\partial x_k}\frac{\partial x_k}{\partial \xi} + \frac{1}{\beta}\frac{\partial \ln |J|}{\partial \xi} \tag{5.83}$$

$$= \mathbf{F} \cdot \frac{\partial x}{\partial \xi} + \frac{1}{\beta}\frac{\partial \ln |J|}{\partial \xi}. \tag{5.84}$$

The two terms R.H.S. of the equation can be interpreted as the mechanical contribution to the force and the change in the volume element from the change in coordinate. If the generalized coordinates are linear functions of the Cartesian coordinates, the second term vanishes [150].

A common example where the volume element change in Equation (5.80) is given by Hénin and Chipot [151]. Consider a 1D RC that is the distance between two particles. The change of coordinates appropriate for this RC is $(x_1, y_1, z_1, x_2, y_2, z_2) \rightarrow (x_m, y_m, z_m, \xi, \theta, \phi)$, where (x_m, y_m, z_m) is the center of the segment joining the two particles and (ξ, θ, ϕ) is the vector between them. For simplicity the coordinates of the other particles in the system are omitted as they do not effect the calculation. The mechanical force between the two particles is simply the difference in the force per particle directed along the vector separating them

$$\mathbf{F}\frac{\partial \mathbf{x}}{\partial \xi} = \frac{1}{2}(\mathbf{F}_2 - \mathbf{F}_1) \cdot \hat{u}_{12}, \tag{5.85}$$

where $\hat{u}_{12}$ is a unit vector. However, the determinant of the Jacobian is dependent on ξ, giving

$$\frac{1}{\beta}\frac{\partial \ln |J|}{\partial \xi} = \frac{2}{\beta \xi}. \tag{5.86}$$

This corresponds to a repulsive pseudoforce between the particles at short distances. However, at sufficiently long distances, the contribution quickly drops below measurable error. The final instantaneous force for this example is

$$F_\xi = \frac{1}{2}(\mathbf{F}_2 - \mathbf{F}_1) \cdot \hat{u}_{12} - \frac{2}{\beta \xi}. \tag{5.87}$$

5.6.2.1 Sampling

The preceding discussion explained how to calculate $\partial A(\xi)/\partial \xi$ for a given $\xi + \Delta\xi$ but not how to sample the range of interest for the RC of choice. Preferably, we would like to uniformly sample our RC to reduce the total error. For TI this is often done by constraining ξ to particular fixed values in a series of runs.

An alternative means to improve sampling is to apply a biasing potential, $U_b(\xi)$ along the RC, giving [143]

$$H'(\mathbf{x}, \mathbf{p}) = H(\mathbf{x}, \mathbf{p}) - U_b(\xi) \tag{5.88}$$

where $H(\mathbf{x}, \mathbf{p})$ is the Hamiltonian and $\mathbf{p}$ is the momenta and Equation (5.78) becomes

$$A(\xi) = -\frac{1}{\beta} \ln \mathcal{P}_\xi + U_b(\xi) + A_0. \tag{5.89}$$

Ideally, for uniform sampling we would like $U_b(\xi) = A(\xi)$ but in practice this is not possible.

Darve, Wilson and Pohorille [143] suggest an iterative scheme where by $U_b(\xi)$ is continuously updated based on results accumulated so far. For an MD simulation, we apply the biasing potential in the form of a force, which, in this case, can be estimated as the current time average force. As this guess is iteratively improved, it is known as an adaptive biasing force (ABF). When $U_b(\xi) = A(\xi)$ we have converged on the solution and the motion along the RC becomes diffusive.

ABF does have some drawbacks. The major theoretical issue is that as U_b is constantly changing, Equation (5.89) is not a true Hamiltonian. However, the Hamiltonian property can, at any time, be recovered by stopping updates to U_b. In practice, this is not an issue. A practical problem with ABF is slow sampling in orthogonal degrees of freedom. This can become the major bottleneck for free energy calculations with this method.

5.6.2.2 Error Calculation

A rough, upper-bound error estimate calculation for ABF is described in Henin and Chipot [151] and Rodriguez-Gomez *et al.* [156]. Briefly, the variance in the average force along the reaction coordinate for a given bin is

$$\sigma^2(\overline{F_\xi}) = \frac{\sum_{i=1}^{p} \sigma^2(\overline{F_{\xi,i}}) + (\overline{F_\xi} - \overline{F_{\xi,i}})^2}{p} \tag{5.90}$$

where σ is the standard deviation, F is the force, ξ denotes the reaction coordinate, p is the total number of bins and $\overline{F_\xi}$ is the average force over the entire RC.

Correlated data reduces the effective number of data points used to calculate an average. For this reason the correlation length, κ, is calculated from a time series of data, X_i. Straatsma *et al.* [157] give κ as

$$\kappa = \sum_{k=1}^{\infty} c_k / c_0 \tag{5.91}$$

where $c_k = \mathrm{cov}(X_i, X_{i+k})$ and $c_0 = \sigma^2(X)$ is then the variance of X. In practice, there is only a finite amount of data and c_k is calculated as

$$c_k = c'_k = \frac{1}{n-k} \sum_{i=1}^{n-k} (X_i - \overline{X})(X_{i+k} - \overline{X}) \tag{5.92}$$

where n is the number data points being considered and

$$\kappa = \sum c'_k / c'_0 \tag{5.93}$$

The final expression for the error in the free energy for a uniform sampling of N points over the entire range of ξ is then

$$\sigma(\Delta A(\xi)) \simeq (\xi_B - \xi_A)\sigma(\overline{F_{\xi,p}}) \frac{\sqrt{1+2\kappa}}{\sqrt{N}} \tag{5.94}$$

Note that κ is calculated for the entire data set without consideration for bins.

Chapter 6

Protein-Solvent Polarization in Myoglobin Hydration[1]

Protein hydration is essential for protein structure and function [73, 74, 78]. Without a minimum amount of water, proteins are not biologically active [73, 81, 158, 159]. Protein hydration has received a great deal of attention in both experiment [81, 158, 159] and simulation [160–163], however, these studies have mostly focused on the structure and dynamics of the protein and water and, to a lesser extent, the polarization of the solvent. Far less attention has been given to protein polarization.

In this study, we focus on the electrostatic polarization of myoglobin and water. The hydration of myoglobin is systematically increased, at ambient temperature and below the glass transition temperature, with the onset of functionality clearly observable. From this, interactions between water and protein can be separated and the role of mutual polarization deduced. Our results show how protein and water influence each other's global structure and this is strongly correlated with protein function.

Section 6.1 introduces myoglobin and the current understanding of its hydration from both simulation and experiment. Section 6.2 details the methodologies we have employed in our simulations and Section 6.3 discusses the results of these simulations.

6.1 Background

6.1.1 Myoglobin

Myoglobin has been one of the most intensely studied proteins. This is partly for historical reasons. Myoglobin was the first protein structure solved with atomic resolution [164] and for this effort Kendrew shared the Noble Prize in Chemistry with Max Perutz [2]. Myoglobin was Kendrew's choice for structure determination because of its relatively small size and it was

[1] A version of this chapter has been published.
Jack A. Tuszynski, Tyler Luchko, Eric J. Carpenter, J. M. Dixon, M. Peyrard, and Yves Engelborghs. Non-Gaussian statistics of the vibrational fluctuations of myoglobin and the thermal fluctuations of myoglobin hydration. Fluctuations and Noise in Biological, Biophysical, and Biomedical Systems II, Proc. of SPIE Vol. 5467 (2004) pp. 1-16

[2] In 1962 both the prizes in Chemistry and Medicine (Crick, Watson and Wilkens) were awarded for work done in Max Perutz' MRC Laboratory of Molecular Biology. Both awards were primarily for the crystal structures of biomolecules (hemoglobin, myoglobin and DNA). That year the Physics prize went to Lev Landau, Literature prize to John Steinbeck and Peace prize to Linus Pauling (his second Nobel).

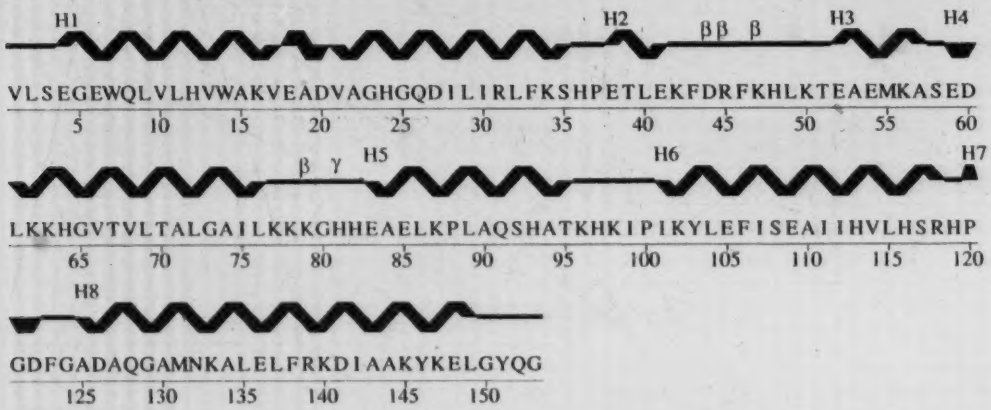


Figure 6.1: Secondary Structure of 1MBC[10]. Eight α helices are shown as coiled ribbons and labeled H1 through H8. Four β-turns are labeled by "β"s and one γ-turn is labeled by a "γ".

easily procured in large amounts, primarily from sperm whales. Because of its size, abundance and the significant amount of knowledge aquired about it, myoglobin continues to be a common choice for both experiment and simulation.

This intense study has also included work on protein hydration under a wide variety of conditions. The amount of work is too vast to summarize here. Some relevent reviews are Phillips and Pettitt [158], Makarov *et al.* [79] and Levy *et al.* [74]. Instead, we focus on the key aspects of myoglobin's structure, function and previous research into protein hydration.

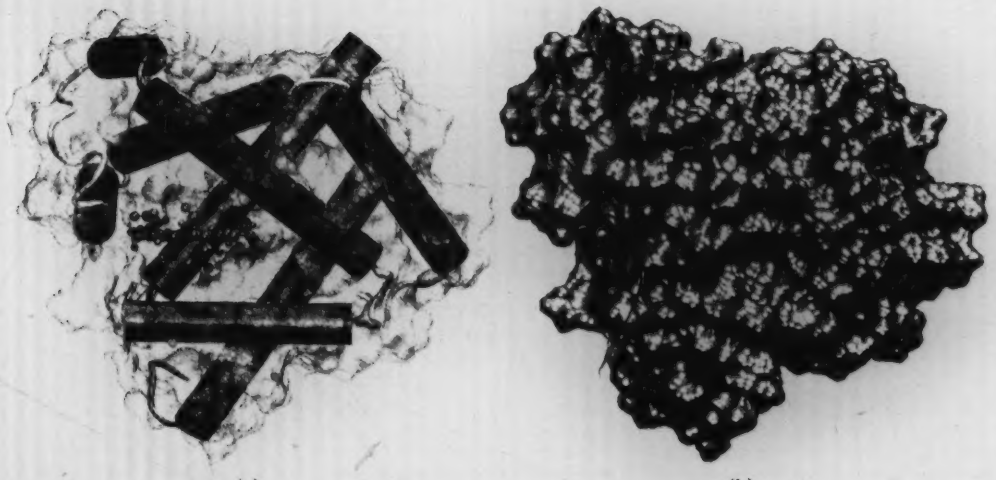
6.1.1.1 Structure and function

Sperm whale carboxy myoglobin (PDB ID 1MBC), is a monomeric, 153 residue protein[8, 165]. Its secondary structure, shown in Figure 6.1, consists of eight α helices (H1-H8), three β-turns and one γ-turn. A prosthetic heme group is covalently bound to residue his-93. In this case, a CO group is covalently bound to the heme iron where O_2 or CO_2 would normally be found. The tertiary structure is shown in Figure 6.2.

The role of myoglobin in the body is to carry oxygen in the muscle, similar to how hemoglobin (see page 7) carries oxygen in the blood. The tertiary structure of myoglobin bears a striking similarity to that of the α subunit of hemoglobin which is also similar to the β subunit. Both hemoglobin and myoglobin are descendant from a common oxygen binding molecule, leghemoglobin. As can be seen in Figure 6.3, the evolution of the globulin family has closely paralleled that of vertebrates. The greatest amount of sequence homology is in residues that stabilize the heme prosthetic group.

6.1.2 Experiment

Experiments have generally been interested in three distinct aspects of protein hydration: the amount of water required for activity (hydration number), the positions of the waters and their dynamics.



(a) (b)

Figure 6.2: Tertiary structure of 1MBC. (a) The tertiary structure is shown inside the water accessible surface of the protein. α helices are red, turns are yellow and random coil is grey. The heme and CO groups are shown as ball and stick models and are coloured grey and cyan respectively. (b) The water accessible surface of the water is shown with hydrophilic residues coloured cyan and hydrophobic residues coloured orange. Images created with VMD [11].

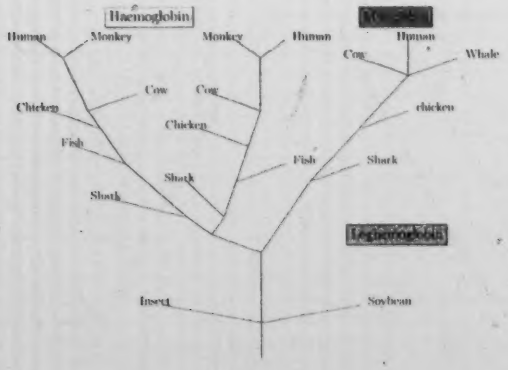


Figure 6.3: Evolutionary tree of the globulin family. (Adapted from [6].)

6.1.2.1 Hydration Number

A protein is considered fully hydrated when the further addition of water does not change the physical properties and serves only to dilute the protein. With such a definition, it is not unexpected that the amount of water required to achieve full hydration, measured in grams of water per grams of protein (h), varies depending on the experimental method.

Differential scanning calorimetry (DSC) is a common tool for measuring the specific heat capacity,C_p, of a sample. In this case, the C_p of a protein powder is measured as a function of h is fully hydrated when C_p stops changing. This is typically around 0.35 h for myoglobin and compares to 0.38 h for lysozyme [158].

Rayleigh scattering of Mössbauer radiation spectroscopy (RSMR) measures changes in the fluctuations and environment of the heme iron. Full hydration in this case occurs at 0.6 h [158]. The higher hydration level may indicate when the heme pocket is hydrated.

A more recently developed method is microwave dielectric spectroscopy [159]. This method has been used to measure the number of waters hydrating myoglobin in native, molten globule and unfolded states with hydration numbers of 0.60 h, 0.64 h and 1.12 h. In this case, the protein is denatured by lowering the pH. This method relies on a number of assumptions; notably, the protein and solvation shells are spherical, the hydration water has the same dielectric properties as bulk and the protein has a constant, uniform dielectric constant of 2.5.

6.1.2.2 Water Positions

The physical location of waters with respect to a protein have been difficult to obtain experimentally. Diffraction and NMR are the two basic methods for this purpose, each with their strengths and weaknesses.

There are three basic diffraction methods that have been used to locate waters around protein: X-ray diffraction, X-ray diffraction with multiwavelength anomalous dispersion (MAD) and neutron diffraction [79]. Both X-ray and neutron diffraction have been used to locate individual waters around proteins. These waters must be tightly bound to be properly fit. However, individual waters are often placed in non-localized electron densities to improve overall fit. This often results in non-overlapping waters for different solutions for the same protein, even for waters on the surface of the protein. Neutron diffraction has been considered more robust as H_2O can be replaced with D_2O to achieve better resolution, thanks to the increased nuclear mass. However, it is unclear if differences in the distributions that result from the two methods are due to differences in the respective methods or differences in hydration properties of D_2O.

MAD has been a marked improvement for obtaining solvent information from protein crystals. The method allows for the direct solution of the electron density of the solvent bulk without flattening procedures that were necessary in previous methods. This gives a 3D solvent density distribution around the protein that can be readily visualized and compared with simulation. For myoglobin, and proteins in general, this method has shown that the first hydration shell typically has 10% greater density than the bulk. It has also been possible to construct so-called proximal radial distribution functions. These map the relative density of the solvent as a function of distance from particular atom types in the solute and have been found to be generally transferable between proteins.

NMR is capable of locating waters about a protein by observing chemical shifts and producing distance restraints between the water hydrogens and the atoms of the protein. This method has been used to locate ordered buried water, water bound to the exterior of protein and, possibly, disordered water in hydrophobic cavities [166]. While this method fails to resolve many of the waters found via diffraction methods it has the benefit of being free of crystal

packing artifacts.

6.1.2.3 Water-Protein Dynamics

The proposed major mechanism for water to impart functionality onto proteins is by influencing its dynamics and vice versa. This may be observed through a variety of methods but here we highlight three.

RSMR has been used to probe the dynamics of the heme iron of myoglobin [81, 158]. This has shown a smooth, non-linear increase in dynamics as full hydration is approached.

NMR is capable of measuring the residence times of waters that associate with the protein [158, 166, 167]. This has been important in demonstrating and quantifying the increased residency times of waters in the first hydration shells and in buried waters.

A recently developed method is tryptophan scanning with a laser probe [168]. Mutants of a target protein (myoglobin in this case) are prepared with single tryptophans in different locations. A laser with a 290 nm wavelength and 90 fs duration is used to excite the tryptophan, causing a sudden change in its dipole moment. The time-dependent emission spectrum is then measured as the side-chain relaxes, which is thought to be dominated by solvent interactions. Regions surrounded by charged side-chains has the slowest relaxation times 100-200 ps. For myoglobin, relaxation times for regions of rigid secondary and tertiary structure were 50-70 ps while loop regions were around 20 ps.

6.1.3 Simulation

Generally speaking, molecular dynamics simulations of protein hydration have sought to reproduce experimental results, providing more detailed observations and, often, postulating explanations or mechanisms.

6.1.3.1 Protein Hydration

The first simulation to systematically hydrate a protein to identify the minimum hydration required for activity was by Steinbach and Brooks [160]. They found only 350 waters were required to hydrate myoglobin, corresponding to a hydration number of 0.35 h. Full hydration was identified as the point when the further addition of water did not change the physical properties measured. However, the short run times involved produce large statistical fluctuations and calculated values were probably not converged. However, this hydration number, though lower than experiment, was still qualitatively correct.

6.1.3.2 Water Positions

The Steinbach and Brooks study also provided insight as to the structure of the solvent around protein. In particular, it showed the patchy nature of the solvation [160]. In fact, a second, weak peak was evident in the protein-water RDFs even at 350 waters. This indicates that it is favourable to hydrate some sites with a second layer before others obtain a first layer.

Fully hydrated protein simulations provide detailed information about the structure and distribution of the solvent around the protein, much of which can be directly compared against experiment [79]. In particular, 3D density distribution functions can be readily calculated and compared with MAD electron density maps, demonstrating the same 10% density increase in the first hydration shell. As with the experimental densities, proximal RDFs can be produced and compare favourably, independent of force field and protein. Furthermore, these simulations locate a vast number of local minima, often separated by less than the radius of a water. This

further emphasizes that the placement of individual water molecules in diffraction data often does not reflect a true solvent distribution.

Combining detailed structural information with empirical force fields offers insights as to the mechanisms behind the structure. For example, Merzel and Smith demonstrated correlations between the orientation of waters and the normal of the protein surface and the various components of the electric field [162, 163].

Extensive work comparing MD simulations to neutron diffration data has been carried out by Tarek and Tobias [169–171]. This work demonstrates that quantitative reproduction of experimental data is possible with appropriately constructed systems. Specifically, systems with free boundary conditions over estimate the fluctuations of the protein atoms, making the potential appear too soft. Periodic boundary conditions with crystalline or random, powder configurations are required to achieve accurate results.

6.1.3.3 Water-Protein Dynamics

MD is uniquely suited for calculating most conceivable dynamical properties of protein and water. Residence times are calculated as the average amount of time a single water molecule lies within a small distance (~ 1 Å) of an identified hydration site [158]. Water residence times for identified hydrations of myoglobin vary from sub-picosecond to 20 ps. Over a variety of proteins, most sites have been identified as having 1-4 ps residence times. A more challenging situation for simulation is buried waters. Since these exchange on time scales that may be $\gg 1$ ns, they are effectively out of the range of MD simulations.

Diffusion rates within the first hydration shell are typically reduced 50% from the bulk [79]. Perhaps more significant, is the anisotropy in diffusion, with diffusion perpendicular to the surface of the protein further restricted, compared to parallel diffusion.

Simulations also have the ability to directly investigate the various contributions to various interactions. This was effectively utilized by applying a restraining potential to all the dihedrals of myoglobin to study their role in hydration [161]. The individual torsions were free to explore their local energy minimum but were prevented from making transitions to other minimum. Myoglobin still displayed the same behaviour under increasing hydration. Rather than influencing protein dynamics by enabling dihedral transitions, it was found that an expansion in the protein backbone occured with hydration, regardless of torsional restraints.

6.2 Methods

Carboxy-myoglobin was simulated at a variety of hydration levels using a methodology similar to that of Steinbach and Brooks [160]. Preparation began with removing a lone SO_4 molecule and 137 water molecules from the crystal structure (PDB ID: 1MBC)[165]. All of the histidines were set to be protonated on the ND1 site only (neutral histidine) giving the molecule a zero net charge. Since the net charge was zero, no ions were added and the system was solvated in a 66 Å diameter sphere of flexible TIP3P [69, 160] water molecules. After molecules overlapping with the protein had been removed 4103 water molecules remained. This was followed by a steepest decent minimization of 100 steps, 1 ps of heating from 0 K to 300 K and 50 ps of equilibration.

After the initial 50 ps of equilibration was performed, waters were removed based on their distance from the protein surface. This resulted in 14 systems of 0, 35, 50, 60, 80, 100, 125, 150, 225, 350, 600, 1000, 1900 and 3830 water molecules, respectively, closely following Steinbach and Brooks [160].

To investigate temperature dependence and qualitative differences above and below the glass transition temperature, each of these systems was then simulated at an equilibrium temperature of 100 K and 300 K, making 24 total simulations. Each simulation was heated for 1 ps to reach its equilibrium temperature and then equilibrated for 150 ps. Production simulations of 300 ps were run for each system and the atomic positions recorded every 0.1 ps.

All simulations were performed with CHARMM [101] 28b using the CHARMM27 forcefield [120]. Constant temperature dynamics simulations used the Berendsen thermostat [115] with a coupling constant of 1.0 ps for equilibration and 5.0 ps for production runs. For heating simple temperature scaling was used. A distance-based cutoff of 28 Å was used for all long range interactions(Lennard-Jones and electrostatic) with a switching function starting at 24 Å [160]. Distance based cutoffs of this length have been shown to capture almost all of the electrostatic energy of the system [172] and should account for polarization effects.

6.3 Results and Discussion

6.3.1 Dipole Moment

For a system of zero net charge the dominant term in the multipole expansion of the electrostatic potential is the dipole moment,

$$V(\mathbf{x}) = \frac{1}{4\pi\epsilon_0}\nabla\left[\frac{q}{r} + \frac{\mathbf{p}\cdot\mathbf{x}}{r^3} + \frac{1}{2}\sum_{i,j} Q_{ij}\frac{x_i x_j}{r^5} + \ldots\right], \tag{6.1}$$

where $\mathbf{p}$ for a system of point charges is

$$\mathbf{p} = \sum_{i=1}^{N} q_i \mathbf{r}_i. \tag{6.2}$$

As such, it is a measure of the shape and strength of an electric field but, unlike the net charge, it may fluctuate. Thus, for an eletrostatically neutral system like hydrated myoglobin it can be used to characterize the protein, water and the system as a whole. Figure 6.4 show that the potential of the myoglobin molecule is overwhelmingly dipolar.

Figure 6.5 show the relationship between the hydration level of the protein and the dipole moment of the protein and surrounding water. Clearly, the addition of water reduces the total dipole moment of the system. Temperature plays a major role as water at 100 K is unable to conform to the electric field of the protein. For the 300 K case the total dipole of the system appears to plateau after hydration of 100 waters. This, however, may be due to the protein's dipole moment at this level of hydration.

6.3.2 Dipole Correlations

Water has a natural hexagonal geometry when in the pure bulk form. In close proximity to protein the network of hydrogen bonds is disrupted as can be seen in Figure 6.6. In close contact individual water molecules conform to the electric field of the protein rather than that of their neighbours. The extent to which water networks are disrupted can be seen by calculating the distance dependence of the correlations between water dipoles and their nearest-neighbors as well as with the electric field produced by the protein.

Nearest-neighbor dipole-dipole correlations are calculated as

$$\langle\cos(\theta_{ij})\rangle_{R_i^{surf}} = \mathbf{p}_i \cdot \mathbf{p}_j \tag{6.3}$$

Figure 6.4: Electrostatic potential of myoglobin. White is negative and gray is positive. The water accessible surface of myoglobin is superimposed with the exposed heme surface colored black. Image created with VMD [11].

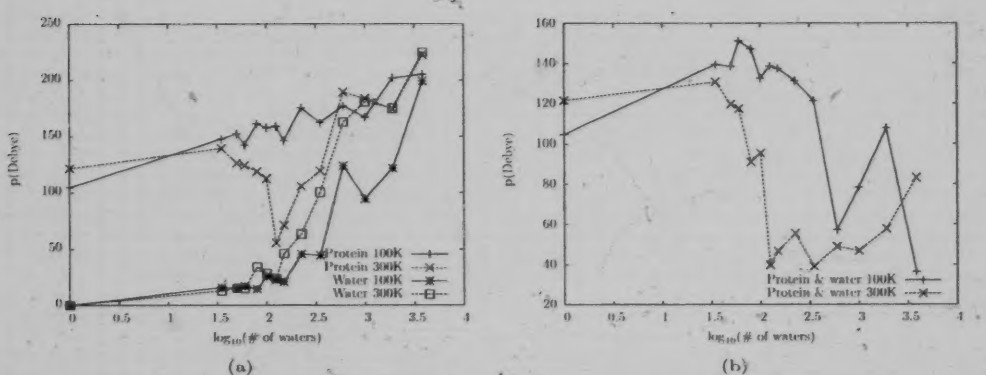


Figure 6.5: Average dipole moment as a function of hydration. Values for protein and water are shown at 100 K (solid lines) and 300 K (dashed lines).

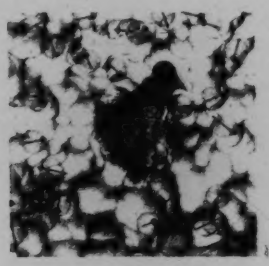
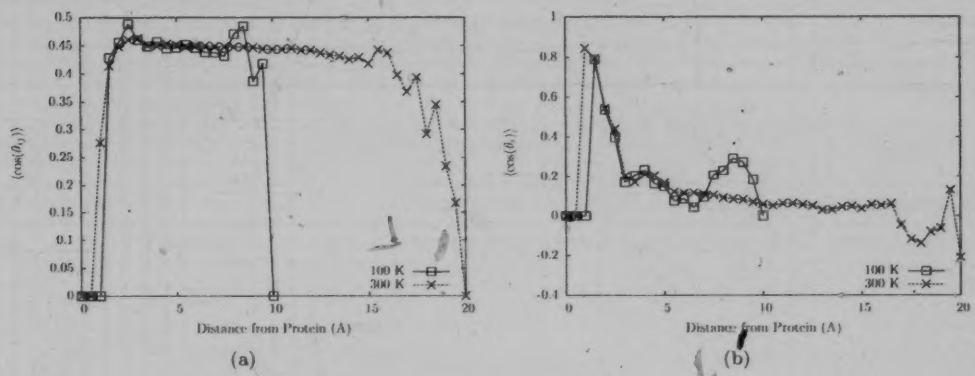


Figure 6.6: Dipole moments of surface water on myoglobin. The water accessible surface of the amino acids is colored white while that of the heme group is colored gray. Individual waters are drawn as spheres with their respective dipole moments draw as vectors inside. Image created with VMD [11].



(a) (b)

Figure 6.7: Nearest neighbour dipole-dipole (a) and dipole-electric field correlations (b) for myoglobin hydrated by 1900 water molecules.

where θ_{ij} is the angle between the dipoles and R_i^{surf} is the distance of the oxygen atom of the i^{th} water molecule from the protein. This is only calculated if the distance between the i and j oxygens is less than 3.5364 Å, corresponding to the minima of the Lennard-Jones potential for the TIP3P water model (see Figure 6.7(a)). At 300 K the water has greater mobility and is able to more easily conform to the potential of the protein near the surface. This is reflected in the more gradual transition between no correlation and a correlation value of about 0.45. The mean correlation of 'bulk' water corresponds to an angle of 64° which is close to what one would expect for a hexagonal lattice. The slow decay of this value demonstrates the range of the protein's influence. Large deviations from this, starting at about 15 Å at 300 K, are distortions due to the water-vacuum interface.

Dipole-electric field correlations directly measure the impact of the electric field of the protein on the orientation of the water molecule. Rather than calculate the angle between the dipoles we calculate the angle between the dipole and the *in vacuo* electric field of protein at the center of the water's oxygen atom

$$\langle \cos(\theta_i) \rangle_{R_i^{\text{surf}}} = \mathbf{p}_i \cdot \mathbf{E}_{\text{protein}}. \tag{6.4}$$

Figure 6.7(b) shows the results of this calculation for myoglobin hydrated by 1900 water

molecules. From this we can see that the higher temperature water is able to get closer to the protein where it is strongly correlated to the protein's electric field. Within about 2.5 Å (which corresponds to about one water layer) of the protein the water is strongly correlated to protein's electric field. Near the edge of the water sphere we see what appears to be surface effects that distort the dynamics of these waters.

6.3.3 Fluctuations, Deviations and Radius of Gyration

As discussed earlier in this chapter, thermal fluctuations are necessary for protein function. How susceptible various parts of a protein are to these fluctuations determines, in large part, what the function of the protein is. We can quantify these fluctuations by looking at the root-mean squared fluctuations (RMSF) and deviations (RMSD). A further indicator of conformational change within the protein is the radius of gyration.

RMSF are defined as

$$\text{RMSF} = \langle \Delta r_i^2 \rangle^{1/2} = \left(\frac{1}{M} \sum_{k=1}^{M} (\mathbf{r}_i(t_k) - \langle \mathbf{r}_i \rangle)^2 \right)^{1/2} \tag{6.5}$$

where $\langle \mathbf{r}_i \rangle$ is average position of atom i over M configurations. This relates to the temperature B-factor, directly measurable from experiment,

$$B_i = \frac{8}{3}\pi^2 \langle \Delta r_i^2 \rangle. \tag{6.6}$$

This gives us a measure of the flexibility of different parts of the protein that can be easily visualized as in Figure 6.8.

Figure 6.8 demonstrates the temperature dependence of the RMSF. Thermal motion is generally confined to the side chains of the amino acids while internal structure, such as α-helices remain rigid even at room temperature. Some parts of the backbone are relatively flexible, such as the N and C-termini.

By averaging the RMSF for different groups of atoms we can see the effect of both temperature and hydration. Hydration does not increase the flexibility of the myoglobin back bone as shown in Figure 6.9(a) and even seems to stabilize it at low temperatures. However, the protein side chains are significantly affected by the addition of water. Figure 6.9(b) shows that after being hydrated by at least 350 waters the flexibility of the side chains is greatly increased to almost three times that of the backbone. At low temperatures this is not the case.

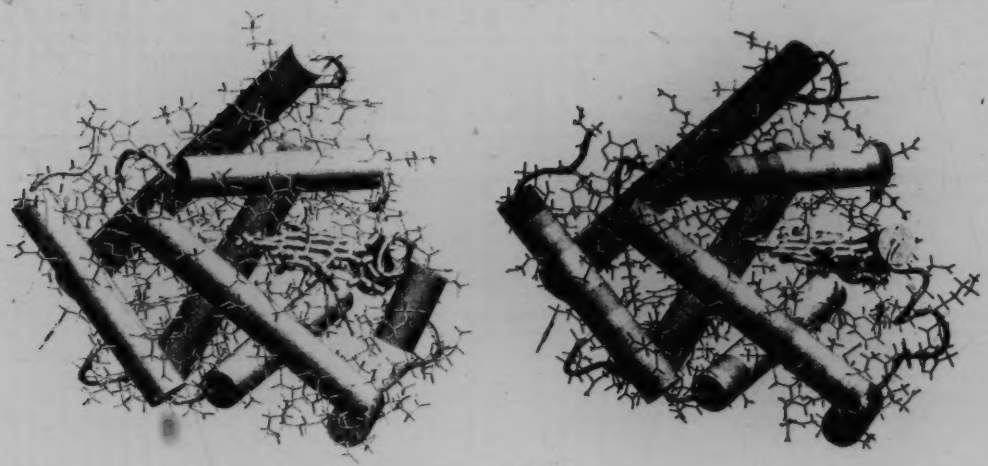
The difference between two structures can be described by the RMSD and is given by

$$\langle \text{RMSD} \rangle = \frac{1}{M} \sum_{i=1}^{M} \left(\frac{1}{N} \sum_{i=1}^{N} (\mathbf{r}_i^A - \mathbf{r}_i^B)^2 \right)^{1/2} \tag{6.7}$$

where N is the number of atoms being compared between structures A and B. Figure 6.10 shows the time averaged RMSD of myoglobin heavy atoms at 100 and 300 K, respectively, compared to the crystal structure. The low temperature structure is closer to the crystal structure, as should be expected, since the crystal structure was imaged at low temperatures. Furthermore, since the high temperature structure has larger thermal fluctuations, it should naturally deviate more significantly. Of interest is that in the 300 K case the deviations plateau after the addition of 350 waters.

Another indicator of conformational change is the radius of gyration. This is defined as

$$R_{\text{gyr}} = \sqrt{\frac{I}{m}} = \sqrt{\frac{\sum m_i r_i^2}{m}}. \tag{6.8}$$



(a) (b)

Figure 6.8: RMSF of myoglobin hydrated by 1900 water molecules at 100 K (a) and 300 K (b). The white-black color scale varies from 0.33 Å to 1.05Å. Images created with VMD [11].

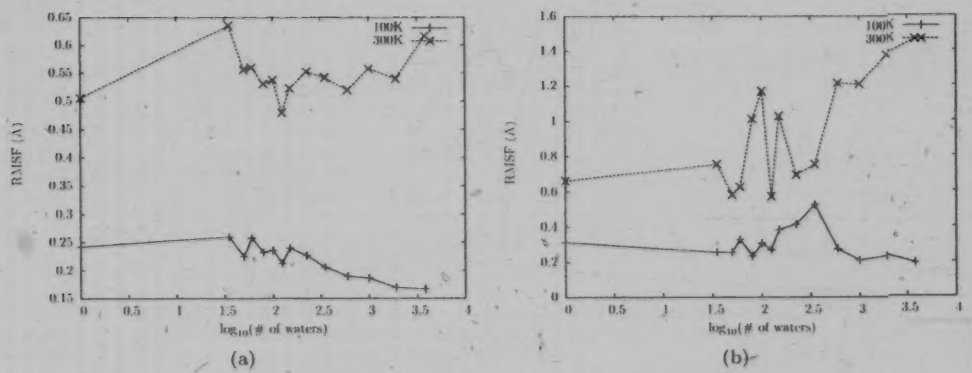


Figure 6.9: Average RMSF of myoglobin backbone (a) and side chains (b).

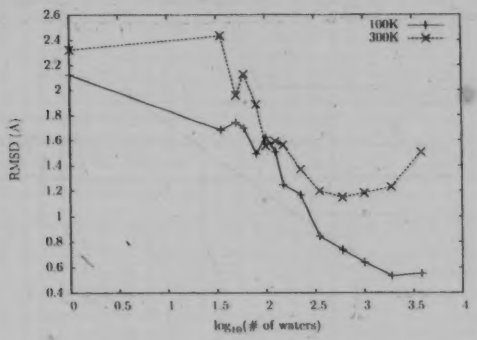


Figure 6.10: Time averaged RMSD of myoglobin heavy atoms.

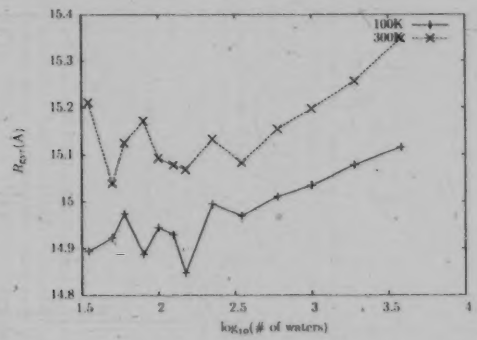


Figure 6.11: Time averaged radius of gyration of myoglobin heavy atoms.

where I is the moment of inertia. This is an indicator of changes in size and shape of the moment of inertia. In the case of myoglobin, being a globular protein, R_{gyr} increases as the protein expands.

Figure 6.11 shows R_{gyr} for the heavy atoms of myoglobin. This value oscillates considerably until hydrated by at least 350 waters. At this point R_{gyr} grows predictably, a strong indication of the minimum hydration being reached at 350 waters/myoglobin. Once again a temperature difference is obvious though volume contractions at low temperature are expected.

6.4 Conclusions

Solvent polarization is known to be an important factor for solvation. However, less emphasis has been placed on solute polarization and its role in protein function.

Our results show, at least qualitatively, that the polarization of myoglobin's electric field is directly connected to the onset of its functionality. As hydration increases, myoglobin's dipole moment first decreases, until 0.35 h is reached, and then increases, lock-step with the increased dipole moment of water. This is evidenced by a plateau in the dipole moment for the system as a whole, starting at 0.35 h.

The effect of proximity to the protein is also demonstrated in the water dipole-dipole and dipole-electric field correlations. Here we see waters within the first hydration shell being

heavily influenced by the protein's electric field and being disrupted from the waters' bulk structure. Interestingly, Figure 6.7 shows evidence for protein water decoupling from each other at low temperatures.

It should be kept in mind, however, that these findings are qualitative in nature. A quantitative study of these properties would require increased sampling and an improved water model. Due to the free boundary conditions, evaporations of waters is inevitable and reduces the practical length of the simulations. This could be overcome with multiple short simulations but would require the preparation of many unique initial configurations. Within the computational resources available for the project, this was not possible.

The TIP3P water model, is an adequate model at 25°C for reproducing water's properties. However, it has not been parameterized for 100 K and can not be expected to properly reproduce any water's behaviour at this temperature. Models, such as TIP4P-Ew [91] and TIP5P [92], have been parameterized over large temperature ranges but this has still not included temperatures as low as 100 K.

Overall, our work shows a strong role in protein function for the polarization of both solvent and solute. The onset of hydration is found to occur at 0.35 h for all measured values, in agreement with previous work [160].

Chapter 7

Molecular Dynamics Calculation of Microtubule Stability

7.1 Introduction

Microtubules (MTs) are ubiquitous, multi-functional organelles found in all eukaryotic cells [7] (Chapter 3). *In vivo*, these structures are highly regular and, almost without exception, consist of 13 protofilament arranged like staves of a barrel. *In vitro*, the internal structure of MTs observed in the cell is still apparent but becomes highly variable and is sensitive to environmental conditions [29, 30]. Detailed background on microtubules is presented in Chapter 3. Experimental [26, 31] and computational [173, 174] efforts have been made to explain the roots of these polymeric structures.

Chrétein and Fuller performed a survey of *in vitro* MT structures and found protofilament numbers ranging from 10 to 16 and typical longitudinal offsets ranging from 7 to 10 Å [26]. It was, however, not possible to differentiate between A and B lattice types. Kikkawa *et al.* [27] demonstrated that B lattices are the dominant type *iv vivo* and *in vitro* using MTs decorated with kinesins and visualized with cryo-electron microscopy (EM). These results agree with other experiments [28].

Sept *et al.* [173] and Drabik *et al.* [174] attempted to explain this preference for the B lattice type via free energy calculations. Sept *et al.* found minima corresponding to A and B lattice types resulting from a solvent accessible surface area (SASA) term. The symmetry between these two minima was broken by electrostatic and solvent polarization terms, calculated with the Poisson-Boltzmann (PB) equation, though they contributed relatively little to the overall free energy. Drabik *et al.* used the 3D-reference interaction site model (3D-RISM) of molecular solvation to calculate the full classical solvation free energy without the need for a SASA approximation. Qualitatively, the two calculations agree, with a global minima corresponding to a B lattice being found. However, the 3D-RISM calculation found no significant minima corresponding to an A lattice. Furthermore, the depth of the potential minima was approximately 200 kcal/mol compared to 25 kcal/mol for the PB-SASA calculation.

Another major topic in MT structure, the proposed conformational change causing depolymerization, has been the domain of experiment so far. The initial evidence for a conformational change has been the, so-called, 'ram's horns' formed by protofilaments upon MT collapse [175]. Subsequent crystal structures two dimer protofilaments stabilized by RB3-SLD [176] and colchicine [15] or vinblastine [16] have demonstrated curvature consistent with that needed for the rams horns. Wang *et al.* [31] used cryo-EM to image tubulin macrotubes

induced with subtilisin proteolysis and divalent ions. Crystal structures of straight and curved tubulins were then docked into the electron densities. The curved tubulin structures were found to be consistent with the GDP tubulin macrotubes while the straight tubulin conformations where consistent with GMPCPP macrotubes. Further, it was proposed that both GTP and GDP tubulin were intrinsically curved.

In this study we seek to determine the microscopic basis for both the MT lattice type and the conformational change observed in MT collapse. To do this we employ all-atom molecular dynamics (MD) with the adaptive biasing force (ABF) method to efficiently calculate the free energy profiles for these interactions. In Section 7.2 we discuss our novel approach to restraining the orientation of an arbitrary, flexible molecule which we use to simplify the reaction coordinates used in our ABF calculations. The details of our simulations are presented in Section 7.3 and the results discussed in Section 7.4 for all three reaction coordinates.

7.2 Orientational Restraint

In practice, it is necessary to reduce the number of degrees of freedom in a system when looking at a reaction coordinate. This is especially true when the reaction coordinate is one-dimensional. Computationally, reducing the degrees of freedom allows the reaction coordinate to be adequately sampled with the resources available. From the analysis point-of-view, the system should be restrained to a physically meaningful pathway.

Often, the simplest approach is to replace explicit solvent in the simulation with an implicit correction term. Unfortunately, this method is not always available and has been known to introduce artifacts in simulations.

Other degrees-of-freedom that can be eliminated are suggested by the system itself and the properties under investigation. In the case of MT structure, we require the tubulin monomers to maintain a MT-like configuration except for specific interactions that are under investigation. This includes not just the relative positions of the monomers but also their orientations.

While it is straight-forward to restrain the center-of-mass of a molecule, restraining the orientation is much more involved. To facilitate this, we have developed a harmonic orientational restraint based off least-squares RMSD fitting of the molecule to a reference structure and quaternion rotations.

7.2.1 Quaternions

Quaternions are widely used for rotations in computer graphics and animations. A description of quaternion algebra can be found from many sources [177–180] and is described below. Karney, in particular, provides a concise review of both the mathematics and various applications to molecular modelling [179].

Quaternions were originally developed in the mid-19$^{\text{th}}$ century as a multi-dimensional extrapolation of complex numbers [177–179]. As the name suggests, quaternions consist of four values and takes the form

$$\mathbf{q} = q_0 + q_1\mathbf{i} + q_2\mathbf{j} + q_3\mathbf{k} \tag{7.1}$$

where q_n are real numbers. $\mathbf{i}$,$\mathbf{j}$ and $\mathbf{k}$ follow associative non-commutative rules for multiplication

$$\begin{gathered}\mathbf{ijk} = -1 \\ \mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = -1 \\ \mathbf{ij} = -\mathbf{ji} = k,\ \mathbf{ik} = -\mathbf{ki} = j,\ \mathbf{jk} = -\mathbf{kj} = i\end{gathered} \tag{7.2}$$

Quaternions may also be represented as an ordered set of four quantities

$$\mathbf{q} = [q_0, q_1, q_2, q_3] \tag{7.3}$$

or as a scalar and vector

$$\mathbf{q} = [q_0, \mathbf{q}] \tag{7.4}$$

where $\mathbf{q} = [q_1, q_2, q_3]$. The conjugate of a quaternion is

$$\mathbf{q}^* = [q_0, -\mathbf{q}] \tag{7.5}$$

Using the scalar-vector notation, multiplication of two quaternions is then

$$\mathbf{qp} = [q_0 p_0 - \mathbf{q} \cdot \mathbf{p}, q_0\mathbf{p} + p_0\mathbf{q} + \mathbf{q} \times \mathbf{p}] \tag{7.6}$$

The squared norm is then given as

$$|\mathbf{q}|^2 = \mathbf{q}\mathbf{q}^* = q_0^2 + q_1^2 + q_2^2 q_3^2 \tag{7.7}$$

and the inverse as

$$\mathbf{q}^{-1} = \mathbf{q}^*/|\mathbf{q}|^2 \tag{7.8}$$

7.2.1.1 Quaternions versus Rotation Matrices

For our purposes, the most important property of quaternions is their ability to perform arbitrary rotations of a vector about an arbitrary axis. To do this we first represent the vector to be rotated in quaternion form

$$\mathbf{v} = [0, v_1, v_2, v_3] \tag{7.9}$$

Given an angle of rotation θ about an axis $\mathbf{a} = [a_1, a_2, a_b]$ we define a unit quaternion

$$\mathbf{q} = [\cos(\theta/2), \sin(\theta/2)\mathbf{a}/|a|] \tag{7.10}$$

The rotated vector, $\mathbf{v}'$ is then

$$= [0, \mathbf{v}'] = \mathbf{q}\mathbf{v}\mathbf{q}^* \tag{7.11}$$

Of course, rotations matrices may be used to perform arbitrary rotations of vectors in 3D-space. In fact, it is possible to change the representation of a given rotation between rotation matrix and quaternion representations. The orthonormal matrix of the unit quaternion

$$\mathbf{q} = q_0 + q_1\mathbf{i} + q_2\mathbf{j} + q_3\mathbf{k} \tag{7.12}$$

is

$$R = \begin{bmatrix} (q_0^2 + q_1^2 + q_2^2 + q_3^2) & 2(q_1q_2 - q_0q_3) & 2(q_1q_3 - q_0q_2) \\ 2(q_2q_1 + q_0q_3) & (q_0^2 - q_1^2 + q_2^2 - q_3^2) & 2(q_1q_2 - q_0q_3) \\ 2(q_3q_1 - q_0q_2) & 2(q_3q_2 - q_0q_3) & (q_0^2 - q_1^2 - q_2^2 + q_3^2) \end{bmatrix} \tag{7.13}$$

The inverse procedure is much more involved. As we do not employ it in this text it the interested reader may consult Horn [180].

Despite the apparent equivalence of the two methods, quaternions have several advantages over rotation matrices. The rotation axis/angle is conceptually straightforward compared to using rotation matrices or Euler angles, particularly in the context of molecular mechanics. Quaternion notation is also more compact, with only four values compared to the nine of rotation matrices. This fact also contributes to the relative speed of applying rotations. As well, the orthonormality of the rotation is easily maintained with quaternions. Finally, quaternions do not suffer from gimbal lock, as do Euler angles, where a rotation in one direction (e.g. 90° in pitch) leads to the other two directions becoming equivalent (roll and yaw). A final benefit of using quaternions is in calculating the optimal rotation to minimize the RMSD between two structures.

7.2.2 Minimizing RMSD

Finding the optimal rotation for the superposition of two equally sized sets of points, $\{\mathbf{r}_{1,i}\}$ and $\{\mathbf{r}_{2,i}\}$, is typically solved by minimizing the root mean squared deviation (RMSD) between the two sets of points[1]. Given that there are M particles, $i = 1 \ldots M$, the RMSD is

$$\mathrm{RMSD}(\mathbf{r}_{1,i}, \mathbf{r}_{2,i}) = \left[\frac{1}{M} \sum_{i=1}^{M} \|\mathbf{r}_{1,i} - \mathrm{R}(\mathbf{r}_{2,i})\|^2 \right]^{1/2} \tag{7.14}$$

where $\mathrm{R}(\mathbf{x})$ is the rotation operator. Horn [180] shows that minimizing the RMSD is equivalent to the problem of maximizing

$$\sum_{i=1}^{M} \mathbf{r}_{1,i} \cdot \mathrm{R}(\mathbf{r}_{2,i}) \tag{7.15}$$

or, in terms of quaternions,

$$\sum_{i=1}^{M} (\mathsf{q}\mathsf{r}_{1,i}\mathsf{q}^*) \cdot \mathsf{r}_{2,i} \tag{7.16}$$

It can be shown that

$$\sum_{i=1}^{M} (\mathsf{q}\mathsf{r}_{1,i}\mathsf{q}^*) \cdot \mathsf{r}_{2,i} = \mathsf{q}^T M \mathsf{q} \tag{7.17}$$

where

$$N = \begin{bmatrix} (S_{xx} + S_{yy} + S_{zz}) & S_{yz} - S_{zy} & S_{zx} - S_{xz} & S_{xy} - S_{yx} \\ S_{yz} - S_{zy} & (S_{xx} - S_{yy} - S_{zz}) & S_{xy} + S_{yx} & S_{zx} + S_{xz} \\ S_{zx} - S_{xz} & S_{xy} + S_{yx} & (-S_{xx} + S_{yy} - S_{zz}) & S_{yz} + S_{zy} \\ S_{xy} - S_{yx} & S_{zx} + S_{xz} & S_{yz} + S_{zy} & (-S_{xx} - S_{yy} + S_{zz}) \end{bmatrix} \tag{7.18}$$

and

$$S_{ab} = \sum_{i=1}^{M} a_{1,i} b_{2,i} \tag{7.19}$$

where a and b may be x, y or z [180]. Also, it can be shown that the unit quaternion that maximizes Equation (7.17) is the eigenvector corresponding to the largest eigenvalue of N. While there does exist an analytic solution to this, in practice it is as fast and easier to use a numerical library, such as Numerical Recipes [181]

7.2.3 Rotational Restoring Force

Once a quaternion has been found that gives the optimal rotation axis and angle (Equation (7.10)), the determination of the restoring torque, τ, is straightforward. The harmonic restraint for rotational energy has the form

$$U_{\mathrm{rot}} = k_{\mathrm{rot}} \theta^2 \tag{7.20}$$

where θ is the angle of rotation from the reference structure and k_{rot} is the coefficient determining the strength of the restoring torque. Thus, the magnitude of the restoring torque itself is

$$|\tau_{\mathrm{rot}}| = 2k_{\mathrm{rot}}\theta. \tag{7.21}$$

[1] Since we are typically interested in the center-of-mass (COM) motion of the of the sets of points, we will assume that the COMs of the two bodies have been translated to the origin. Neither the mass weighting nor the rotation point of the origin is necessary. We simply must agree on a common point of rotation.

with the direction being the axis of rotation given by the quaternion. However, the standard Verlet equations and, therefore, the MD packages that implement them, work only in forces. Thus, we need to compute the instantaneous force for each atom to reproduce the restoring torque.

There are an number of properties that we would like the applied torque/forces to have. Obviously, the total applied torque should be equal to the restoring torque:

$$\sum_{i}^{N} \boldsymbol{\tau}_i = \boldsymbol{\tau}_{\text{rot}}. \tag{7.22}$$

Furthermore, to prevent distortion of the structure, we should have an equal angular acceleration, α, for each atom, which is also the angular acceleration of the entire object:

$$\boldsymbol{\alpha}_{\text{rot}} = \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = \cdots = \boldsymbol{\alpha}_N. \tag{7.23}$$

The angular acceleration on each particle is given by torque applied and its moment of inertia

$$\boldsymbol{\alpha}_i = \frac{\boldsymbol{\tau}_i}{I_i} = \frac{\boldsymbol{\tau}_i}{(r_i^2 m_i)} \tag{7.24}$$

where I is the moment of inertia, r the distance from the COM and m is the mass of the particle. Combining this with Equations (7.23) gives

$$\frac{\boldsymbol{\tau}_{\text{rot}}}{I_{\text{tot}}} = \boldsymbol{\alpha}_{\text{rot}} = \boldsymbol{\alpha}_i = \frac{\boldsymbol{\tau}_i}{(r_i^2 m_i)}. \tag{7.25}$$

So,

$$\boldsymbol{\tau}_i = \frac{r_i^2 m_i \boldsymbol{\tau}_{\text{rot}}}{I_{\text{tot}}}. \tag{7.26}$$

Once the restoring torque for each particle has been assigned, forces must be computed. Forces are related to torques through the equation

$$\boldsymbol{\tau}_i = \mathbf{r}_i \times \mathbf{F}_i. \tag{7.27}$$

However, there is no inverse operation for the cross product. To see this, we explicitly state the equation for each component of the torque,

$$\begin{aligned} -r_{zi}F_{yi} + r_{yi}F_{zi} &= \tau_{xi} \\ r_{zi}F_{xi} - r_{xi}F_{zi} &= \tau_{yi} \\ -r_{yi}F_{xi} + r_{xi}F_{yi} &= \tau_{zi}. \end{aligned} \tag{7.28}$$

This is a system of three equations and three unknowns (F_{xi}, F_{yi} and F_{zi}), the solution of which would provide the inverse cross product and determine $\mathbf{F}_i$. However, this is an inconsistent set of equations with no solution unless certain conditions are applied.

One set of conditions that does provided a solution is projecting the system onto a 2D-plane, normal to the torque. Since an arbitrary rotation can be applied to the coordinate system, this is easily expressed as $\tau_{xi} = \tau_{yi} = 0$, $\tau_{zi} = |\tau_i|$ and $\tau_z = 0$. The general solution in this case is

$$F_{xi} = F_{xi},\ F_{yi} = \frac{\tau_{yi}F_{xi} + |\tau_i|}{r_{xi}},\ F_{zi} = 0 \tag{7.29}$$

where F_{xi} is a free variable. The actual torque applied, $\tilde{\tau}_i$, can be determined by substituting this solution in to Equation (7.28), giving

$$\begin{aligned} \tilde{\tau}_{xi} &= -\tau_{zi}\frac{\tau_{yi}F_{xi} + \tau_{zi}}{\tau_{xi}} \\ \tilde{\tau}_{yi} &= \tau_{zi}F_{xi} \\ \tilde{\tau}_{zi} &= \tau_{zi}. \end{aligned} \tag{7.30}$$

Our choice of F_{xi} determines the distribution of error between $\tilde{\tau}_{xi}$ and $\tilde{\tau}_{yi}$.

An alternative method, similar to that used in NAMD for restraints relative to a rotating reference [104, 182], uses the fact that the applied force should be perpendicular to both the position vector and the desired torque to be applied. This can be calculated using a cross product as

$$\mathbf{F}_i = \boldsymbol{\tau}_i \times \mathbf{r}_i. \tag{7.31}$$

Once again, the actual torque applied can be calculated as

$$\begin{aligned} \tilde{\tau}_{xi} &= \tau_{xi}(r_{yi}^2 + r_{zi}^2) - \tau_{yi}r_{xi}r_{yi} - \tau_{zi}r_{xi}r_{zi} \\ \tilde{\tau}_{yi} &= \tau_{yi}(r_{zi}^2 + r_{xi}^2) - \tau_{xi}r_{xi}r_{yi} - \tau_{zi}r_{yi}r_{zi} \\ \tilde{\tau}_{zi} &= \tau_{zi}(r_{xi}^2 + r_{yi}^2) - \tau_{xi}r_{xi}r_{zi} - \tau_{yi}r_{yi}r_{zi} \end{aligned} \tag{7.32}$$

Using the same coordinate system as the previous method (τ is aligned with the z-axis), this method over estimates the τ_z while the error in the x and y components is always negative. The error in this method results from the fact that while the applied force is perpendicular to the desired torque and the position vector, the position vector and desired torque are not perpendicular to each other. This method does have the desirable trait that no net force is applied to the object.

Both methods suggested here are not capable of calculating the exact forces to apply an arbitrary torque to a collection of particles. A reasonable estimate is produced, though the angular acceleration depend on the location of the particle relative to the center of mass.

For the calculations presented here, the second, cross product, method was used as it does not impart a net force. No deleterious artifacts were observed.

7.2.4 Implementation

Horn's method for optimized superposition of two sets of particles [180], and our restraint method based on this, were implemented as a TCL extension written in C for NAMD [104]. By implementing the methods in C instead of TCL computational efficiencies were gained. This is particularly important as TCL extensions to NAMD are not run in parallel and represent a computational bottleneck. This method is able to restrain the 3D orientation a set of particles to a reference system. Alternatively, this restraint may be performed along a specific axis. Care should be used in this case as a rotation of 180° about an axis perpendicular to the restraining axis is still possible. If this occurs, the method becomes unstable.

7.3 Simulation Protocol

Tubulin-tubulin PMFs were calculated along three different reaction coordinates (Figure 7.1). Inter-protofilament interactions were calculated for lateral separation and longitudinal displacement. Three intra-protofilament interactions were calculated for bending at the N-site and E-site interfaces, corresponding to the proposed conformational change leading to MT collapse.

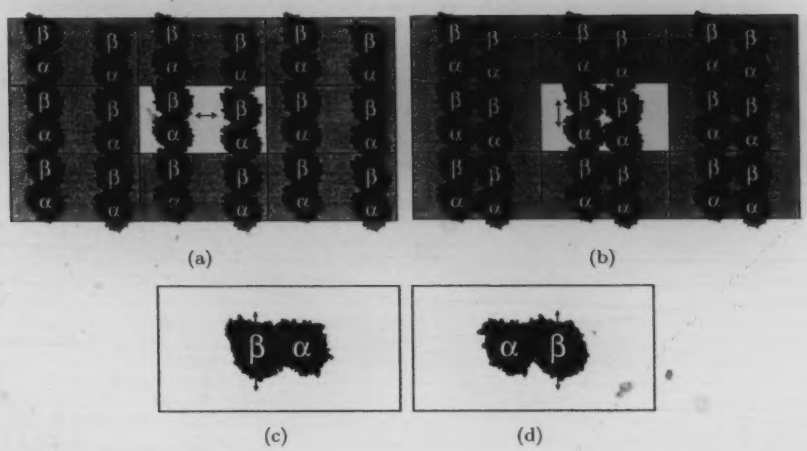


Figure 7.1: Reaction coordinates for MT inter-protofilament interactions in lateral (a) and longitudinal directions (b) and intra-protofilament interactions (c) and (d).

The PMF for the E-site interfaces was calculated in both GTP · MG and GDP states. In total, five PMFs were calculated.

All five PMFs were constructed from the same initial model of tubulin, based on the crystal structure of Nettles *et al.* (PDB ID: 1TVK) [18]. The crystal structure failed to resolve both N-terminal methionines for each monomer, residues 35 to 60 of α-tubulin and both C-terminal tails (CTTs) (residues beyond 439 for α-tubulin and 427 for β-tubulin). Note that for the missing CTTs this is one more residue for α-tubulin than would be present after subtilisin proteolysis and five fewer residues for β-tubulin [55]. All missing residues, except for the CTTs, were replaced with MODELLER [183] and colchicine bound tubulin crystal structure (PDB ID: 1SA0) [15]. As α-tubulin residues 38 to 46 were also missing in this structure, 1000 loop variations were constructed with MODELLER and the lowest energy structure was used for all subsequent calculations. As only incomplete nucleotide structures were present, their positions and structures were taken from another the refined tubulin-taxol crystal structure (PDB ID: 1JFF) [14].

Protonation states for all titratable residues were determined with PROPKA [184] and manual inspection of each group. Three PROPKA calculations were performed for a single tubulin dimer in alternate configurations: $\alpha\beta$ with the E-site exposed, $\alpha\beta$ with the N-site exposed and $\alpha\beta$ with the N-site exposed and GTP · MG at the E-site. From this procedure, His-37β and His-267β were assigned a doubly protonated state while all other histidines were protonated on the γ-nitrogen only. Also Glu-429α and Asp-251β were observed to have significant pKa shifts and were assigned neutral, protonated states. These protonation states were applied to all configurations. The final patched and protonated dimer was then rotated 64° about the x-axis to conform to the MT model of Li *et al.* [22].

For the two systems in a $\beta\alpha$ configuration to investigate the E-site interface, the β-tubulin monomer was translated 81.2 Å along the x-axis. A periodic box of 205 Å × 110 Å × 140 Å was used for all simulations.

Inter-protofilament calculations consisted of two dimers aligned to conform to the MT model of Li. A periodic box with dimensions 81.2 Å × 75 Å × 125 Å was used to simulate infinite protofilaments along the x-axis. In total, four initial systems were created with different

offsets between the two dimers along the x-axis. These offsets were 9.32 Å (corresponding to a B-lattice), 30 Å, 50 Å and 70 Å.

All seven initial systems were solvated with a replicated box of pre-equilibrated TIP3P water molecules [69]. Waters with the most negative electrostatic potential were iteratively replaced with K^+ counter ions until neutrality was reached. Subsequently, Na^-K^+ ions replaced random waters until salt concentration of 0.1 M was achieved. The Amberff03 force field [122] was applied using the TLEAP module of Amber [5].

The systems were then minimized for 10000 conjugate gradient steps using NAMD [104]. This was followed by heating to 310 K over 20 ps with 50 kcal/mol harmonic restraints applied to all protein atoms. These parameters were maintained for an additional 50 ps to equilibrate the water. The harmonic restraints on the protein were then linear decreased over 500 ps. All systems were then simulated with center-of-mass harmonic restraints for the positions and orientations of the monomers. The lateral distance between protofilaments was allowed to fluctuate freely. Equilibration continued until the potential energy and the root mean squared deviations of coordinates for the initial conditions were both observed to plateau. The simulation time required for this was typically 5 ns for the intra-protofilament systems and 10 to 15 ns for the inter-protofilament systems.

ABF production runs for the protofilament bending simulations used RC bins of $\Delta x =$ 0.35 Å and a force constant of 100 kcal/mol/Å at the RC boundaries. The initial RC was bounded by a -1.75 Å and 1.75 Å displacement along the y-axis initially and expanded to encompass the experimentally determined angle range as the simulation progressed. A center of mass position and orientational restraint was placed on the α monomer to ensure a maximum deviation from the initial orientation of no more than 0.1°. No restraints were place on the β monomer. The displacement along the RC was the difference in position of the two center-of-masses along the y-axis. Thus, the equilibrium position is not necessarily zero.

The inter-protofilament, longitudinal offset RC was divided into eight pieces with each of equilibration simulation spawning two production runs. Each simulation covered approximately 10 Å of the total RC with 1 Å bins. The only overlap was the 9-10 Å bin. An orientation restraint on each protofilament maintained the MT model orientation.

A single simulation was used for the lateral offset RC. The longitudinal offset was restrained to 9.32 Å, as in the equilibration run, and used a 0.2 Å bin size. Harmonic restraints prevented the lateral offset or the displacement along the z-axis from occurring. An orientation restraint on each protofilament maintained the MT model orientation.

7.4 Results and Discussion

7.4.1 Protofilament Offset

The PMF for protofilament-protolfilament interactions as a function of longitudinal offset is given in Figure 7.2. The lateral separation for the two protofilaments is given in Figure 7.3.

7.4.1.1 Convergence

Immediately obvious from Figures 7.2 and 7.3 is that the calculation has not converged. As this is a periodic system, we expect the free energy difference between the two endpoints to be zero but find a 10 kcal/mol difference instead. Similarly, Figure 7.3 shows the distribution of the lateral protofilament COM separation, which should be continuous throughout, including the endpoints. While small artifacts do occur at the boundaries of some of the individual runs, clearly, the largest discrepancy is between the two endpoints. The large difference observed here is primarily due to the protofilaments separating early in the production phase for the

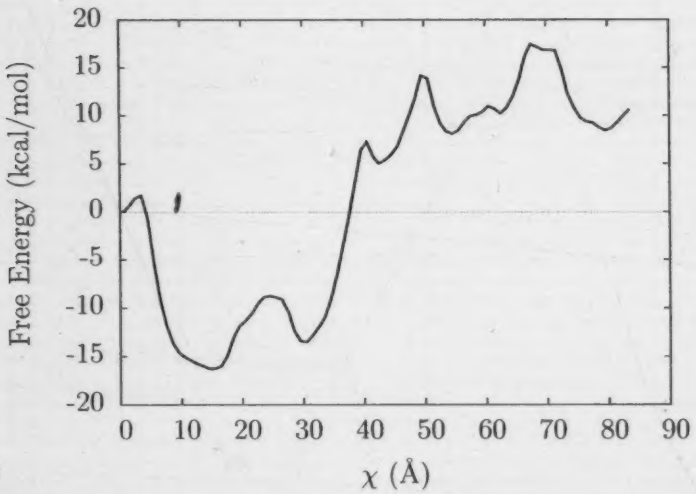


Figure 7.2: Free energy profile of protofilament interactions along a longitudinal offset.

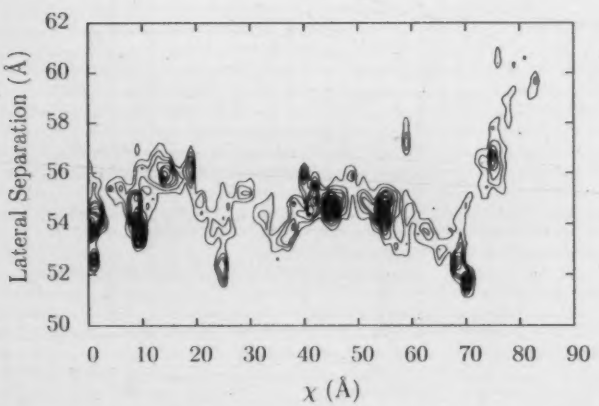


Figure 7.3: Protofilament-protofilament separation vs. longitudinal offset.

72-84 Å region of the RC. While the protofilaments did return to a small separation for all points on the RC, a large fraction of the sampling occurred with a larger separation. The overall incomplete sampling is confirmed by examination of the trajectory and sampling of χ_{long} (data not shown).

The extensive simulation time, together with the lack of convergence, suggests slow relaxation in degrees of freedom orthogonal to χ_{long}. These degrees of freedom clearly impact the physical behaviour of our system, particularly along the reaction coordinate.

Simple visual inspection of the all-atom trajectory clearly implicates the M-loop as the main impediment to convergence. Figure 7.4 shows the distortion of the M-loop as the protofilament offset is increased. The M- and N-loops have previously been identified as two of the main contacts between protofilaments in an MT. The distortion, rather than sliding, of the M-loop is consistent with the M-loop having a strong interaction with the H1-S2 loop on the adjacent protofilament. We can also speculate that the M-loop may impart both flexibility and strength to the MT. As the MT is distorted, lateral interactions via the M-loop undergo significant distortion before breaking, providing more elasticity and resilience to shear and bending forces.

However, this distortion prevents the convergence of the simulation by introducing a history dependence. Figure 7.4(b) shows the distortion of the M-loop with $\chi_{\text{long}} = 19.9$ Å and equilibrated at $\chi_{\text{long}} = 9.34$ Å while Figure 7.4(c) is a conformation at $\chi_{\text{long}} = 20.2$ Å and equilibrated at $\chi_{\text{long}} = 30$ Å. Sampling of the similar M-loop conformations for adjacent displacements is only likely to occur after an extremely long time.

The N-loop, however, displays almost no distortion. Though the two loops display little sequence identity (about 40%), the reason for the N-loops apparent rigidity is likely the S9-S10 loop [14]. Figure 7.4 shows this loop in both monomers. Both steric and electrostatic interactions allow the S9-S10 loop to stabilize the N-loop.

This also provides insights to the functional mechanism of taxiods and epothilones. The site of the 'missing' eight residues of the S9-S10 loop in β-tubulin is also the binding site of the taxiod and epthilone classes of MT stabilizing agents. Considerable experimental evidence has shown that paclitaxil stabilizes the M-loop. In crystallography experiments, the M-loop has been resolved if, and only if, paclitaxil or epothilone have been present in the binding site [13–15, 17, 18]. Hydrogen/deuterium exchange (HDX) coupled to liquid chromatography-electrospray ionization mass spectroscopy has also shown a stabilization of the M-loop [185]. There is little doubt that these drugs stabilize both MTs and the M-loop. The effect of these drugs on MT mechanical properties is a source of some controversy however.

7.4.1.2 Flexural Rigidity

MT flexural rigidity has been experimentally measured in one of two ways, through observation of thermally induced distortions or mechanical manipulation. Howard and colleagues have used video analysis, coupled with Fourier mode decomposition, to calculate flexural rigidity based on observed changes in shape due to thermal fluctuations [186, 187]. A 1.2 fold increase in rigidity relative to GTP-capped MTs was observed at 37°. A temperature dependence was also observed, with Taxol stabilized MTs being 1.7 times stiffer at 37° compared to 25°. Vale *et al.* did not measure the rigidity but simply the curvature of MTs with and without Taxol and guanylyl α,β-methylenediphosphate (GMPCPP) (a very slowly hydrolyzed GTP analog) [188]. While a small decrease in means curvature was observed for GTP and GMPCPP MTs when Taxol was added, it was only statistically significant for GMPCPP MTs and very small compared to the effect of GMPCPP itself. Felgner *et al.* [189] used optical tweezers to bend and oscillate MTs with and without Taxol and MAPs. In both these experiments Taxol was found to reduce rigidity by almost four fold. Kawaguchi *et al.* [190] only measured the rigidity of Taxol stabilized MTs but used both thermal and kinesin buckling methods. The

Figure 7.4: Distortion of tubulin M- and N-loops at longitudinal offsets of (a) 9.2 Å, (b) 19.9 Å and (c) 20.2 Å. (a) and (b) were equilibrated at $\chi_{\text{long}} = 9.34$ Å and (c) was equilibrated at $\chi_{\text{long}} = 30$ Å. α- and β-tubulin are shown as red and blue ribbons while the minimized crystal structure at $\chi_{\text{long}} = 9.34$ Å is shown in grey. Residues in the M- and N-loops, and residues with 5 Å on the adjacent protofilament, are shown as sticks with the following colouring: basic:blue, acidic:red, polar:green, non-polar:white. Images created with VMD [11].

(a) (b)

Figure 7.5: Taxol (PDB ID: 1JFF) [14] (a) and epothilone (PDB ID: 1TVK) [17] (b) binding sites. The compounds are coloured by atom type: carbon:cyan, nitrogen:blue, oxygen:red and sulphur:yellow. Residues within 5 Å of each compound are coloured as in Figure 7.4. Images created with VMD [11].

results are in good agreement with previous measurements except for those of the Howard group. It is noted that for all experiments the Howard group consistently used the longest MTs by a factor of two or more. This may have enabled them to capture longer modes not observable in the shorter MTs used by other groups.

Comparing our PMF to those of Drabik *et al.* and Sept *et al.*, it is tempting to think of these previous models being Taxol stabilized as the M-loop is effectively rigid. However, there are other differences in these models. The Drabik PMF used protofilaments arranged in a sheet configuration, giving somewhat different contacts than in an MT. Sept's PMF only used an explicit atom representation for electrostatic interactions. Entropic contributions, including steric and hydrophobic interactions, were included only as a simple surface area term, without specificity for the atom types involved in the interaction. Thus, sequence differences in the N- and M-loops are not accounted for. Finally, none of the M-loop conformations used in either study are the result of typical MT contacts. M-loops in the Sept model were taken directly from the refined crystal structure of tubulin [14] where the contacts for the M-loop are known to be different than in an MT. In the Drabik model various M-loop configuration were created via simulated annealing but these are representative of free tubulin.

The precise effect of the interaction between Taxol and the M-loop is not known. Figure 7.5 shows tubulin with Taxol and epothilone. Only part of the void left by the S9-S10 loop is filled by either of these compounds. Furthermore, there are no strong electrostatic interactions made. This suggests that the effect of these compounds is to push the M-loop out, improving lateral contacts but not necessarily impeding the shear flexibility observed in our simulations.

The H1-S2 loop is another flexible part of tubulin's structure that plays an important role in lateral contacts, with a similar deformation to that seen in the M-loop. In the distorted lattice shown in Figure 7.4(b) and (c), the α-tubulin H1-S2 loop maintains contact with the M-loop if approached from a low χ_{long} value but interacts with the N-loop if approached from a high value. The complete structure of this loop for α-tubulin has not been observed in any crystal structure and contributes to the elasticity along χ_{long} as the M-loop. Once again, this structure is frozen in both the Drabik and Sept calculations.

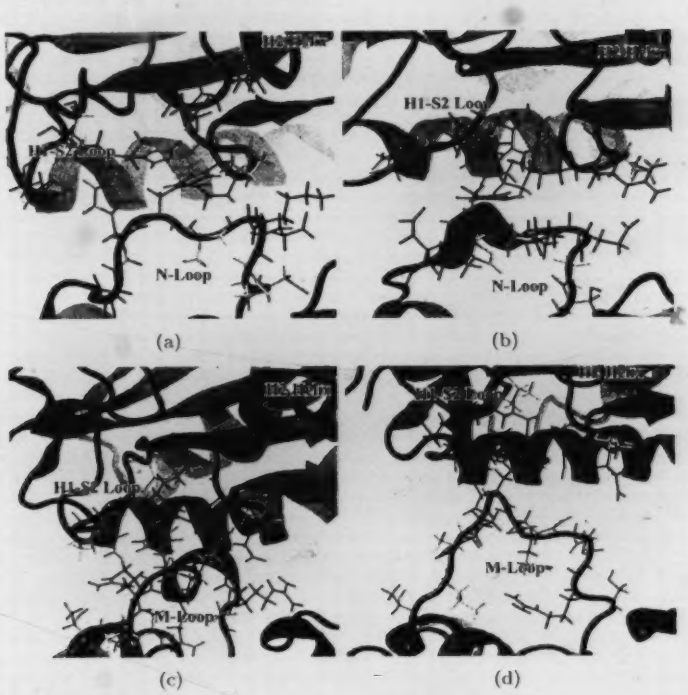


Figure 7.6: N- and M-loop interactions for A- and B-type lattices. (a) N-loop interactions for A-type lattice. (b) N-loop interactions for B-type lattice. (c) M-loop interactions for A-type lattice. (d) M-loop interactions for B-type lattice. Colouring as in Figure 7.4. Images created with VMD [11].

7.4.1.3 Lattice Type

While not fully converged, the PMF shows a clear preference for the B-type MT lattice. While this is qualitatively in agreement with both the Sept and Drabik calculations, the presence of only one significant minima is in closest agreement with the Drabik PMF. In particular, the lack of a minima corresponding to the A-type lattice is the most prominent difference between the calculations.

The likely root of this difference is the sequence difference between the M- and N-loops. The largest contribution to the free energy in Sept's calculation is from the surface area term that does not differentiate between atom or residue types. The similar conformations of the two loops (recall the M-loop is in a Taxol stabilized conformation) gives the same basic potential well for the A-type and B-type longitudinal offsets. Figure 7.6 shows the observed interactions for the N- and M-loops for A- and B-type offsets.

7.4.2 Protofilament Separation

Protofilaments were free to explore their lateral separation while calculating the PMF for longitudinal separation. This is explicitly explored with an RC, χ_{lat}, laterally separating the protofilaments, restrained to a B-type lattice offset. Figure 7.7 shows a minimum at 55.1 Å. Given that this calculation only covers the extent of the M-loop's contact range, we can not

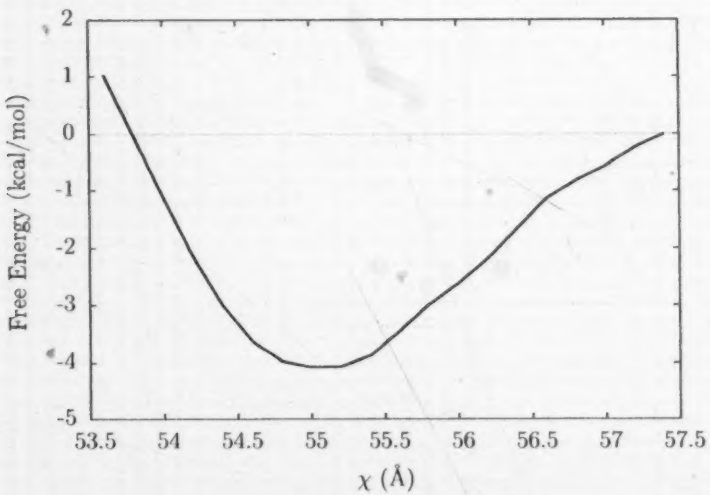


Figure 7.7: Free energy profile of protofilament interactions along a lateral separation.

estimate the depth of this well and can only compare to experiment in terms of the observed protofilament separation. The Li *et al.* [22] model gives this to be 54 Å.

The only other attempt to calculate this PMF from theory has been by Drabik *et al.* [174]. It is difficult to compare these numbers, due to the extreme difference in magnitude. However, qualitatively, we see that they place the minima at 56-57 Å depending on the conformer of the M-loop used. In fact, one conformer has a minimum at 59 Å. This, again, demonstrates the importance of M-loop flexibility. Figure 7.8 shows the M- and N-loops at three points along the RC: $\chi_{\text{lat}} = 53.4$, 55.1 and 57.4 Å. While there is some stretching in the N-loop, it is not as significant as in the M-loop.

Figure 7.8: M- and N-loops at different lateral protofilament separations. (a) M-loop at $\chi_{\text{lat}} =$ 53.5 (light blue), 55.1 (blue) and 57.5 Å (purple). (b) N-loop at $\chi_{\text{lat}} = 53.5$ (pink), 55.1 (red) and 57.5 Å (orange).

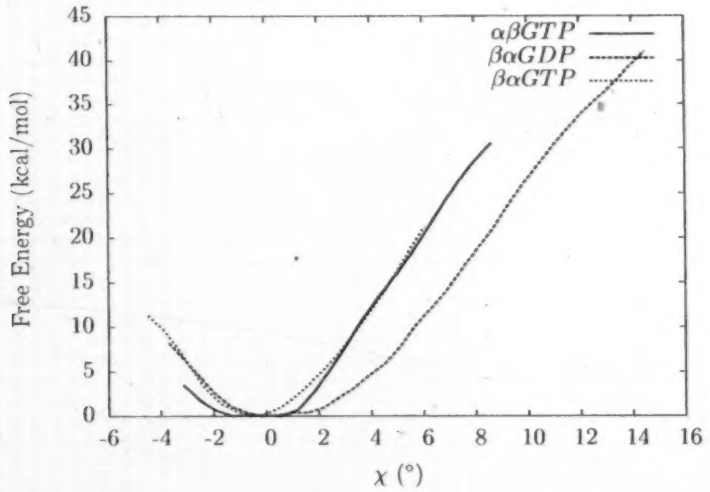


Figure 7.9: Free energy profile of protofilament bending. The PMF is corrected to be relative to the rotation angle from the crystal structure.

7.4.3 Protofilament Bending

Protofilament bending was examined with an RC based on monomer displacement along the MT radial direction. This is consistent with the proposed conformational change in tubulin and the observed structures (rams horns and rings) that are a product of MT collapse [31, 175]. The bending angle between monomers is often used as a measure of this. A similar measurement is the change in orientation angle relative to the straight conformation, in this case, taken from tubulin protofilament sheets. We observe a linear correlation between radial displacement and the orientational angle (data not shown). Thus, using a simple linear fit, we convert between the radial displacement and orientational angle. The resulting PMF is shown in Figure 7.9.

Convergence is, once again, an issue for this calculation. In this case, each PMF was calculated with one simulation and there is no apparent obstruction to the calculation. Therefore, it should readily converge given enough simulation time.

There are several important differences in the three PMFs. The N-site interface, $\alpha\beta$GTP, shows a flat bottomed well with a steep wall, giving some flexibility around the equilibrium angle but a strong penalty for large deviations. $\beta\alpha$GTP is quite similar to $\alpha\beta$GTP but does not have the same flat bottom to the PMF, suggesting that the E-site interface with GTP may be slightly stiffer than the N-site interface. The $\beta\alpha$GDP PMF has the same flat bottom of the N-site, but slightly shifted, and a softer penalty for larger deviations. Overall, it is the most relaxed configuration. Nowhere is there observed a second minima or any other evidence of a conformational change.

This overall stiffening of the interface in the presence of GTP is consistent with experimental findings of Vale *et al.*, who found that GMPCPP induced a significant reduction in the curvature of MTs. As GTP cannot hydrolyze in our simulations, it is effectively the same as GMPCPP.

While there is no conformational change, there does appear to be a structural basis to the softening of the PMF. The S3-H3 loop has been proposed as an analog to the switch-II region of classical GTPases [191, 192] and a similar structure in FtsZ [193]. This switch-II region of these classical GTPases does not undergo a conformational change. Rather, the region relaxes

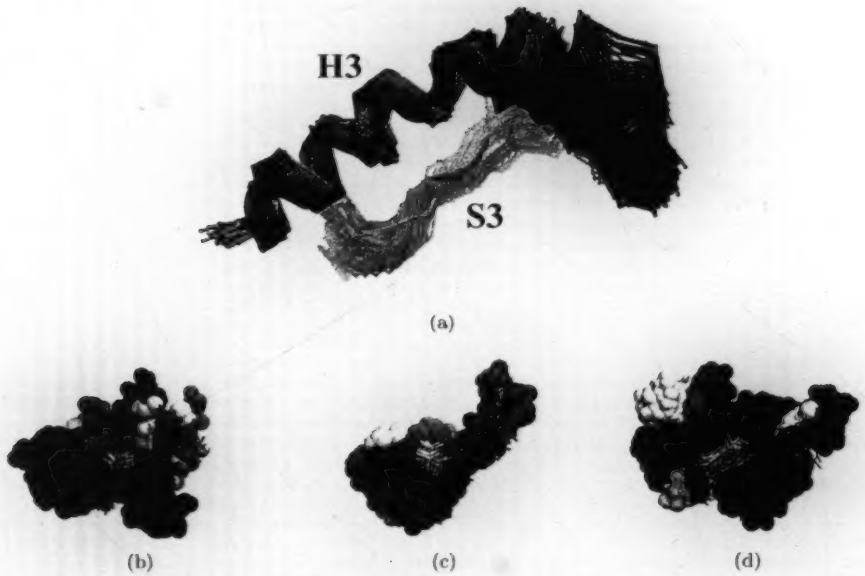


Figure 7.10: S3-H3 interface. (a) Superposition of one structure per nanosecond for simulations of GDP-β-tubulin (blue), GTP-β-tubulin (light blue) and GTP-α-tubulin (red) with the crystal structures of straight α-tubulin (grey), β-tubulin (black) (PDB ID 1TVK) and curved β-tubulin (PDB ID 1SA0). Superposition is based on the minimization of the RMSD for the H3-helix and S3-sheet. The interface for the simulation structures with residues within 5 Å on the adjacent monomer are shown for GDP-β-tubulin (b), GTP-β-tubulin (c) and GTP-α-tubulin (d). Residues on the adjacent monomer are illustrated with a space-filling model while for the S3-H3 loop only the bonds between heavy atoms are drawn. Colouring for individual residue is as in Figure 7.4. Images created with VMD [11].

and become less structured. In our simulations the loss of the terminal phosphate and Mg^{2+} results in a closer association of the S3-H3 loop to the β-monomer. Figure 7.10(a) shows how the S3-H3 loop is pushed out by the presence of GTP relative to the conformations seen for GDP β-tubulin and α-tubulin, both in simulation and the crystal structures. Figures 7.10(b)-(d) highlight the differences in the interfaces resulting from the different nucleotide-monomer combinations. Two important observations are made. First, we note that the E-site interface in Figures 7.10(b) and (c) is dominated by polar sidechain interactions. That is, hydrogen bonds provide the majority of the positive interaction energy. For the N-site interface in Figure 7.10(d) four salt bridges provide for a much stronger bond. The other observation is the change in the interacting residues for the E-site for the different nucleotides. When GTP is present, there are well defined interactions with a small number of polar side-chains on α-tubulin. The presence of GDP causes the S3-H3 loop to interact with non-polar residues and a single, like-charge interaction.

These three variations of the interface have important consequences when comparing and interpreting the simulation results against experimental structures. In particular, conformations with with RB3-SLD, colchicine and vinblastine have been suggested as the natural bent conformation of the tubulin-GDP complex [15, 31]. Figure 7.11 compares the structure at the largest bending angle for the E-site interface to the experimentally determined bent con-

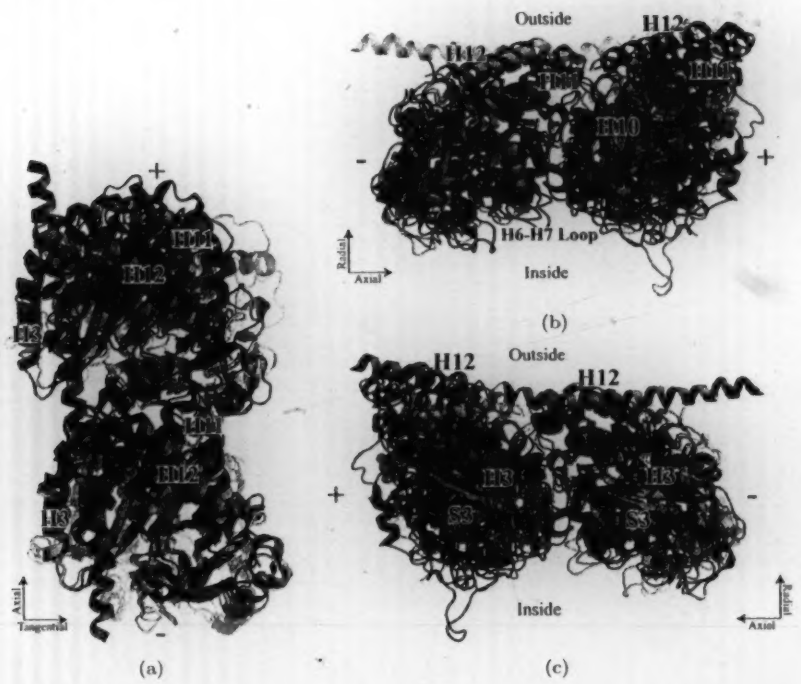


Figure 7.11: E-site conformational bending with views from outside the MT (a) and tangential to the MT surface ((b) and (c)). α- and β-tubulin from the simulation are colour red and blue respectively with the nucleotides shown as space filling models. The straight crystal structure of tubulin is shown in black (PDB ID: 1TVK) and the curved in yellow (PDB ID: 1SA0). The relevant RB3-SLD fragment from the curved structure is shown in orange. Images created with VMD [11].

formation. Two exceptional differences are to be noted, out of plane bending and lack of a H6-H7 loop shift. The simulated structure naturally distorts tangentially to the MT surface as well as the forced perpendicular motion (Figure 7.11(a)). This is contrary to the 1SA0 crystal structure where the bend is in the pure radial direction. The main conformational difference between the straight and 1SA0 protofilament structures is a shift in the H6 and H7 helices of β-tubulin, relative to the C-terminal domain, to maintain contact with the H10 helix of α-tubulin (Figure 7.11(b)). This does not occur in our simulations. Rather, the H6-H7 loop maintains contact with the H10 helix by the α monomer shifting and bending at an angle in the plane tangent to the MT surface. This is accompanied by a gap forming between the S3-H3 loop and the α monomer (Figure 7.11(c)). The likely cause of this discrepancy is the stabilizing RB3-SLD polypeptide present in the crystal structure. The α-helix of RB3-SLD runs the entire length of the two dimer complex and binds, in part, at the C-terminal end of the S3-H3 loop. This effectively acts to bind together the S3-H3 loop with the α monomer and to prevent curvature along the MT surface tangent plane.

The N-site interface demonstrates the importance of the four salt bridges present on the α-tubulin S3-H3 loop. The same views of the bent configurations as in Figure 7.11 show that the simulated structure does not bend in the MT surface tangent plane bend at the E-site

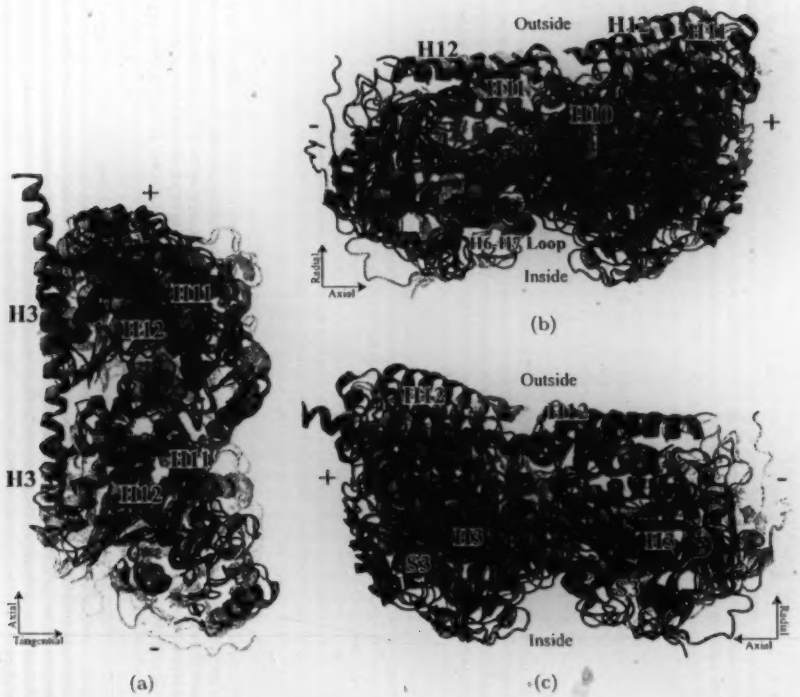


Figure 7.12: N-site conformational bending with views from outside the MT (a) and tangential to the MT surface ((b) and (c)). Colouring is as in Figure 7.12 with the addition of colchicine shown as a space-filling model coloured green. Images created with VMD [11].

(Figure 7.12(a)), even with a radial displacement that is larger than in experiment. However, the H10 helix on β-tubulin maintains contact with the H6-H7 loop and does not distort as in the 1SA0 structure (Figure 7.12(b)). Also, only slight movement in the S3-H3 loop and the corresponding binding site on the β tubulin is observed but a tight contact is maintained.

Together, these suggest a model in which a GTP hydrolysis causes the β S3-H3 loop to associate more tightly with the nucleotide, weakening the bond to α-tubulin and softening the potential interaction. This allows a bending at the E-site at an angle to the MT axial and radial directions. Though the α-tubulin shows a nearly identical S3-H3 loop conformation to GDP β-tubulin, the four salt-bridges the loop forms with β-tubulin prevent bending at the N-site interface.

Not explained are the near identical shifts in the H6-H7 loop and H10 helix for both α- and β-tubulin in the presence of colchicine for the 1SA0 crystal structure. Although colchicine is in a good position to move the H10 helix in β-tubulin, it is not present at the E-site and does not explain the shift in α-tubulin. Also, since we do not observe a similar shift in H10 for β-tubulin at the N-site, it appears that this is not a response of the protein to being bent radially outward. Thus, the shift is not caused by RB3-SLD. It is possible that the shifts are coincidently similar and are cause by colchicine and the RB3-SLD fragments respectively for the for the N- and E-site. However, this seems improbable.

7.5 Conclusions

PMFs were calculated for three basic tubulin-tubulin interactions. A longitudinal offset between adjacent protofilaments was used to investigate the basis of MT lattice structure while a lateral offset probed the separation of protofilaments. Bending of individual dimers explored the proposed conformational change, thought to be responsible for MT depolymerization.

Lateral interactions have demonstrated the role of the M-loop as a dynamic, flexible tether between the protofilaments. These loops are able to stretch out to maintain contacts between protofilaments when either a shear force (longitudinal offset) or depolymerizing force (protofilament separation) is applied. In fact, the M-loops are able to maintain their original contacts even over shear deformations of at least 1 nm.

It is clear that the M- and N-loops play an important role in MT flexibility. However, it is unclear what effect stabilizing these loops, particularly the M-loop with Taxol, has on MT mechanical properties. Conflicting experimental data further obscures the problem. It is possible that Taxol stabilizes the M-loop such that lateral contacts are better maintained, but does not impede the (sliding) flexibility of the loops when shear forces are applied. Repeating part of the above calculation, with a limited RC (only in the vicinity of the B-lattice offset) and Taxol or an extended S9-S10 loop in the M-loop pocket, could offer important insights.

The H1-S2 and S2-H3 loops display similar flexibility to the M- and N-loops. As with the M-loop, stabilizing either of these loops may enhance MT stability. This may offer the possibility of a novel binding site for a completely new MT stabilizing agent. The lack of a known natural toxin that targets this part of tubulin suggests it to be unlikely.

While the flexibility of the M-loop is an important finding, the nature of this stretching had a negative impact on the convergence of the PMFs. Because of the hysteresis introduced by the stretching, dividing the RC into multiple windows means that the final results, even when fully converged, do not smoothly connect at the boundaries. To calculate a PMF comparable to the Sept and Drabik calculations using (ABF) MD, a 2D RC should be used to simultaneously sample longitudinal offsets between the protofilaments and longitudinal positions of the M-loop relative to its monomer. The results here suggest this would still be an enormously expensive calculation and the flexibility of the H1-S2 and S2-H3 loops may still pose a problem for convergence and sampling. If studying shear forces is the primary goal, this can still be accomplished used a limited longitudinal RC, for example, the calculation presented here limited to a ± 10 Å offset from the A- or B-lattice configurations.

There is considerable structural experimental evidence mounting suggesting that free tubulin has a bent native conformation, or at least that a straight and bent conformation exists. Our calculations exploring this conformational change suggest this may not be the case. The picture we find is that the potential minima is softened but the bent conformation is still strongly penalized relative to the straight one. Of course, the PMF calculated here is not for a protofilament in an MT but a free dimer. In a full MT there may be other, entropic forces involved. Regardless, while it is plausible that there is no preferred bent conformation, such a model must still explain experimental observations. This includes not only depolyermization but measured depolymerization forces [194].

Whether or not there is a native bent conformation of tubulin, the notion of the protofilaments folding radially outward from the MT may also be overly idealized. Rather there may be a tangential component as well. This is supported by the observation that ram's horns form spirals as well as rings or tubes when MTs depolymerize [175, 176, 195]. Important in this model is a small shift in the position of the S3-H3 helix. This effectively loosens the E-site interface, allowing more bend both radially from the MT and tangential to its surface. In fact, it is possible that the direction of the bending is sufficiently orthogonal to the RC calculated here that a significant change has been lost in the broadened bottom of the free energy well.

This can be further explored using PMF calculations. However, as the exact RC is not known, this is best done with a 2D RC. Along with an RC perpendicular to the MT surface, an RC tangent to the MT surface should also be included. If a bending direction does exist, it will appear as an extended basin in the 2D RC.

Chapter 8

Conformational Analysis of the Carboxy-Terminal Tails of Human β-Tubulin Isotypes[1]

8.1 Introduction

Microtubules (MTs) are hollow cylinders constructed from linear chains of the protein tubulin. Tubulin is highly conserved throughout the entire eukaryotic kingdom and the two main classes, α/β are expressed from multiple genes, producing several isotypes with seemingly identical functions [196–198]. At the cellular level, it has been hypothesized that MT stability is regulated by subtle variations observed between α and β isotypes [199–201]. There is approximately 80-95% sequence identity between isotypes, however the extreme carboxy-terminal tails (CTTs) exhibit considerable differences, having only 50-60% identity in this region [42, 43]. Early phylogenetic comparisons of the vertebrate β-tubulin families identified the CTTs, along with an internal variable domain as the primary isotype defining features of the β-tubulin protein [202].

The importance of the CTTs has been demonstrated through their removal using limited proteolytic cleavage by substilisin. Following cleavage, the critical tubulin concentration required for polymerization was found to be approximately 50 times lower than that for intact tubulin [56]. This rate was shown to decrease further through the addition of MT stabilizing compounds such as paclitaxel [54]. Proteolyzed tubulin also exhibits altered protofilament bending, resulting in the formation of sheets, bundles of twisted filaments, rings, unstructured aggregates, or MTs with reversed polarity [56, 203]. Finally, the presence of the highly charged CTTs is thought to obstruct tubulin/tubulin interactions, regulating the rate and conformation of MT assembly through unfavorable electrostatic interactions [204]. This hypothesis is supported by observation that divalent cations, or the substitution of acidic residues results in an increased rate of MT assembly and decreased rate of MT disassembly [205, 206].

In addition to tubulin interactions within the MT itself, the functional stability of MTs *in vivo* has been shown to depend on interactions with microtubule-associated proteins (MAPs) through interactions with the CTTs. For example, the correct assembly of the kinetochore and

[1] A version of this chapter has been published.
T. Luchko, J.T. Huzil, M. Stepanova, and J. Tuszynski. Conformational analysis of the carboxy-terminal tails of human beta-tubulin isotypes. Biophys. J., 94:1971-1982, Mar 2008.

the integrity of the mitotic spindle is dependent on the ability of the Dam1 complex to bind the CTTs of β-tubulin [207]. Interactions between the CTTs and MAP2 or MAP tau have also been shown to result in the stabilization of MT structures *in vivo* [208, 209]. Interactions between kinesin and MTs may also be significantly modulated through direct contacts with the CTT [60] or through the indirect modulation of kinesins ability to bind ATP [59]. Recent evidence also suggests that the CTTs may play a role in apoptosis, as interactions between MTs and the pro-apoptotic and anti-apoptotic members of the Bcl-2 family are also affected by the presence of the CTT regions of α and β-tubulin [210], an observation that may offer a plausible explanation for the previously observed interactions between Bcl-2 and paclitaxel [211].

While the presence of intact CTTs is seemingly essential for proper MT assembly and function, little is known about their structure and how this may impact interactions with other proteins. Several early studies have examined possible conformations of the CTTs using NMR and CD spectroscopy and demonstrated that there was inherent disorder within the CTTs [45, 46]. Although a region of increased helicity towards the amino-terminus was identified, a finding that was subsequently confirmed by electron crystallographic analysis, this has provided little additional information about CTT structure as the last 10 residues of α-tubulin and the last 18 residues of β-tubulin were not visualized due to lack of density [13]. A possible explanation for this lack of density is the non-homogenous presence of tubulin isotypes in the MT preparations used in the crystallographic analysis. However, homogenous samples of $\alpha\beta$II and $\alpha\beta$III tubulin failed to improve the quality of density maps and Nogales et al. (1998) concluded that their lack of resolution was indeed due to disorder in this region.

Fortunately, molecular modeling provides us with the unique ability to examine conformations of the CTTs at a level of detail experimental analysis is unable to yet provide. However, even modeling has its limitations in this regard, where a persistent problem with low temperature protein-folding simulations is that of obtaining adequate sampling. This problem exists because simulations generally become trapped in one of a large number of local energy minima. Several generalized ensemble algorithms, based on non-Boltzmann probability weight factors, are capable of overcoming this problem by introducing a random walk in energy space [133]. However, it is often not a trivial matter to determine the non-Boltzmann weight factors and, for this reason, we have chosen replica exchange molecular dynamics (REMD) as a method to examine possible conformations of the tubulin CTTs [140, 212]. REMD utilizes a large number of parallel simulations at different temperatures, with exchanges between trajectories attempted periodically using Metropolis criteria. All replicas of the system then perform a random walk through temperature space and as a consequence, also through energy space. A replica may therefore overcome an energy barrier through exchange with replicas at higher temperatures. This allows configurations to be sampled at a given temperature on time scales not otherwise possible while still maintaining a thermodynamically consistent ensemble of configurations. Here, we discuss the creation of nine models of human β-tubulin CTT peptides using REMD and their analysis for relative conformational flexibility within the ensemble.

8.2 Methods and Materials

Peptide Construction and System Configuration. Based on our previous examination of β-tubulin isotypes, we chose a set of nine peptides corresponding to the consensus CTT sequences of βI (GI:34222261), βII (GI:68299771 and GI:42476191), βIII (GI:50592995), βIVa (GI:21361321), βIVb (GI:68051719), βV (GI:14210535), βVI (GI:41152077), βVII (TUBB4Q) (GI:55770867) and βVIII (TUBB8) (GI:42558278) (Table 8.1) [42]. Here, accession numbers correspond to Entrez Nucleotide expressed mRNA sequence IDs contained only within the human genome. Using the crystallographic structure of tubulin (PDB ID: 1JFF) as a guide,

Isotype	Sequence	Length	Charge
I	DATAEEEED-FGEEAEEEA--------	18	-12
II	DATADEQGE-FEEEEGEDEA-------	19	-12
III	DATAEEEGEMYEDDEEESEAQGPK---	24	-12
IVa	DATAEEGEF--EEEAEEEVA-------	18	-11
IVb	DATAEEEGE-FEEEAEEEVA-------	19	-12
V	DATANDGEEAFEDEEEEIDG-------	20	-12
VI	DAKAVLEED--EEVTEEAEMEPEDKGH	25	-11
VII	DATAEGGGV----------	9	-3
VIII	DATAEEEED--EEYAEEEVA-------	18	-12

Table 8.1: ClustalW multiple sequence alignment of the CTT sequences. The length and net charge at pH 7.0 was determined for each peptide, including the C-terminal cap.

we selected the conserved DA[TK]A motif that is positioned at the extreme end of helix H12 as the initiation point for the construction of all the peptides [213]. This position marks the beginning of the structurally undefined region of the CTT in the structure and the domain examined using NMR and CD spectroscopy [45, 46]. All nine CTTs were prepared identically in an extended conformation, having the N-terminal charge neutralized by capping with an acetyl group using the PyMol v0.99 residue and fragment builder facility [214]. The conventional protonation state at pH 7.0 was used for all residues, with attention paid to His, which was protonated on the ϵ nitrogen throughout.

8.2.1 Parameterization and Model Preparation

All calculations utilized the Amber99 force-field [121, 215], which was applied using PDB2GMX in GROMACS 3.3 [216]. Each peptide was placed in a rhombic-dodecahedron unit cell consisting of approximately 2800 TIP3P waters [92]. Sodium and chlorine counter-ions were added to the most energetically favorable locations (as determined using GENION) such that the net charge of the system was neutralized and a final ion concentration of 100 mM was established. For all MD and REMD calculations rigid bonds were maintained using the LINCS algorithm for the peptide and SETTLE for waters. Particle Mesh Ewald (PME) was used for electrostatic interactions with a cutoff of 0.8 nm and a Lennard-Jones cutoff of 1.0 nm. Constant pressure was maintained by allowing the box size to fluctuate isotropically. Each system was minimized using a Low-Memory BFGS minimizer in GMXRUN until machine precision was achieved. This was followed by 20 ps of heating and 1.1 ns of equilibration. REMD. To achieve a transition probability of 0.3, 43 target temperatures between 273 and 382 K were selected. Each replica was then heated/cooled to its target temperature over 50 ps and simulated without exchange for 500 ps, followed by 500 ps of REMD. Production REMD runs consisted of 10 ns of dynamics for each replica with exchanges attempted every 5 ps. Energy and conformational snapshots were saved every 1 and 10 ps, respectively. For each isotype, this produced an ensemble of 43,000 conformations and an aggregate simulation time of 430 ns over all temperatures.

8.2.2 Cluster Analysis

A cluster analysis of the resulting REMD conformations was used to determine preferred conformations and relative populations of each peptide. The GROMOS clustering algorithm [217], implemented in `G_CLUSTER`, was used for this purpose. All Cα atoms were RMSD fit to the ex-

tended starting structure and an RMSD cutoff was set for the Cα atoms and, for each structure from the ensemble, all structures within this cutoff were assigned as neighbors. The structure with the most neighbors was then identified as the center of a cluster and was removed from the pool with all its neighbors. This process was repeated until all structures had been assigned to a cluster. The RMSD cutoff was chosen such that the largest cluster contained 50% plus 1 of the structures.

8.2.3 Principal Component Analysis

As MD simulations tend to produce immense quantities of data, Principal Component Analysis (PCA) is a powerful mathematical tool used to detect correlations in MD trajectories [218, 219]. To perform PCA, all of the conformations from the REMD ensemble were RMSD least squares fit to the reference structures, effectively removing all rotations and translations. Then, for non-mass-weighted PCA, the covariance matrix can be calculated as:

$$\sigma_{ij} = \langle (r_i - \langle r_i \rangle)(r_j - \langle r_j \rangle) \rangle \tag{8.1}$$

where $r_1 \ldots r_{3N}$ are the Cartesian coordinates of the N atoms used in the analysis and denote the ensemble average. The resulting matrix can then be diagonalized and the resulting $3N$-dimensional eigenvectors, $\mathbf{v}_i$, are organized in descending order of eigenvalues, λ_i. Eigenvalues represent the variance along their associated eigenvector and the larger the eigenvalue the more significant the correlated motion.

A principal component analysis of all nine isotypes of β-tubulin was performed at 311 K using the positions of all Cα atoms. To mimic the CTTs bound at their N-terminus, the backbone atoms of the first three amino acids of the ensemble were RMSD fit to the reference structure for the isotype at 311 K, producing an average structure (Figure 8.1). This was essential for the physiological relevance of the PCA calculation and the considerable motion of the CTTs. Eigenvalues of different isotypes cannot be directly compared, as different numbers of atoms were used in the covariant analysis. This can be overcome by normalizing the eigenvalues by the number of atoms used in the analysis:

$$\sigma_i = \sqrt{\frac{\lambda_i}{N}}, \tag{8.2}$$

where σ_i is the root normalized eigenvalue representing the standard deviation along the i^{th} eigenvector.

8.2.4 Sequence Alignments

Sequence alignments of CTTs were performed with the default values using the European Bioinformatics Institute Clustal server (http://www.ebi.ac.uk/clustalw/), with the exception that the extended gap penalty was increased to a value of 0.5 (Table 8.1).

8.2.5 Motif Identification

Structural motifs within the CTT ensemble were identified by performing pair-wise alignment of each CTT sequence using the ClustalW alignment facility in MacVector (MacVector, Inc, North Carolina, USA). Motifs were identified as those sequences containing at least four consecutive identical residues. A structural similarity search was then performed using only sequence runs of four or more identical or similar residues. Structural similarity was determined by

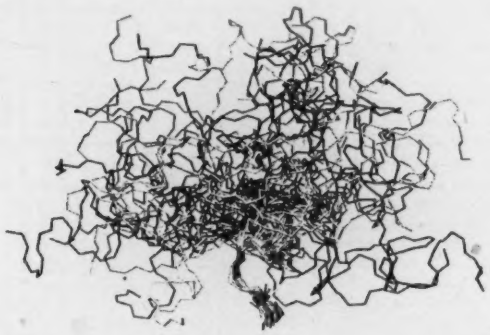


Figure 8.1: Illustration of reference and average structures used for PCA. The N, C and Cα atoms from the three N-terminal residues of each conformation from the βI ensemble were RMSD fit to a representative structure backbone. The backbone atoms of 50 randomly selected conformers of the ensemble are shown in black, gray and white to delineate conformers in no specific order. The first four residues of each peptide are shown in yellow. The average structure, used for subsequent PCA calculations, is shown in light blue. Image created with VMD [11].

calculating the ρ_{sc} between all pairs of conformations from the ensembles of each isotype to create a 1001 × 1001 matrix. Here, ρ is defined as

$$\rho(A,B) = \frac{2 \cdot \mathrm{RMSD}(A,B)}{\sqrt{R^2_{gyr}(A) + R^2_{gyr}(B) - RMSD^2(A,B)}} \tag{8.3}$$

where A and B are structures with an equal number of points, R_{gyr} is the radius of gyration, and RMSD stands for the root mean squared deviation of the two structures. To obtain ρ_{sc}, we translate the centroid of each structure to the origin and rotate it such that the principle axes of inertia lie on the coordinate axes. Then, for both structures, we satisfy the condition that $R_{gyr} = 1$ by scaling each axis independently such that each contributes equally to the radius of gyration. The ρ_{sc} provides a length independent measure of conformational similarity of two structures [220]. Structures with a ρ_{sc} of 0 are considered identical, those with a factor of 2 are considered maximally dissimilar. Mirror images have a value of $\sqrt{2}$ while structures with a factor of 0.3-0.5 indicate visual similarity. The fraction of conformational pairs with a ρ_{sc} of less than 0.3 is the fraction of time that regions within the two isotypes are structurally similar. Regions that showed higher than 50% ρ_{sc} similarity between isotypes are listed in Table 8.2.

8.3 Results

8.3.1 Amino Acid Composition of CTT Consensus Sequences

The absolute length of each human βtubulin CTT ranges from 9 to 25 residues, however their typical length falls between 18 and 20 residues (Table 8.1). The exceptions are βVII, which contains only nine residues, and βIII/βVI which have 24 and 25 residues, respectively. While the overall amino acid composition of each CTT is quite similar, there are some notable factors to consider. When we examine all nine human CTT sequences in aggregate, the residue that occurs most frequently is Glu (42%). The occurrence of all other residues drops significantly

CK2 Motif

Isotype	Range	Sequence	Isotype	Range	Sequence	Similarity
III	1-9	DATAEEEGE	IVb	1-9	DATAEEEGE	72.56
II	1-8	DATADEQG	III	1-8	DATAEEEG	76.64
II	1-8	DATADEQG	IVb	1-8	DATAEEEG	62.97
II	1-5	DATAD	VII	1-5	DATAE	61.40
IVa	1-5	DATAE	VII	1-5	DATAE	64.00
II	1-4	DATA	V	1-4	DATA	64.10
IVb	1-4	DATA	V	1-4	DATA	62.34
V	1-4	DATA	VI	1-4	DAKA	51.26
V	1-4	DATA	VII	1-4	DATA	67.19
V	1-4	DATA	VIII	1-4	DATA	52.90
VI	1-4	DATA	VII	1-4	DATA	52.70

MAP2 Motif

Isotype	Range	Sequence	Isotype	Range	Sequence	Similarity
II	10-14	FEEEE	III	11-15	YEDDE	60.89
II	10-14	FEEEE	V	11-15	FEDEE	71.61
IVa	11-14	EEAE	VI	14-17	EEAE	71.87
I	12-15	EEAE	VI	14-17	EEAE	78.71
IVb	12-15	EEAE	VI	14-17	EEAE	80.93
I	15-18	EEEA	III	15-18	EEES	67.62
IVa	14-17	EEEV	V	15-18	EEEI	69.12
IVb	15-18	EEEV	V	15-18	EEEI	85.56
V	15-18	EEEI	VIII	14-17	EEEV	55.65

Table 8.2: Sequence motif identification. Pairwise alignments of each CTT sequence identified several similar regions between peptides. Structural similarity between sequences was determined by calculating the ρ_{sc} between all pairs of conformations. Those pairs with high conformational and sequence similarity were used to characterize the motif conformations for the CK-2 and MAP2 domains (Figure 8.4).

from this value, Ala (18%), Asp (12%), Gly (8%), Thr (5%), Val (4%) and Phe (3%). The βVIII CTT contains no Gly, the βIVa/b CTTs have a proportionally low Asp content and finally the βVII CTT has a reduced Glu content, which may simply be a result of its length. Interestingly, the remainder of the amino acids occur exclusively in the βIII (Lys, Met, Pro, Gln, Ser, and Tyr), βV (Ile, Asn) and βVI (His, Lys, Leu, and Pro) CTTS at frequencies of 1% or less. Finally, Cys, Trp and Arg are absent from all the CTT sequences.

8.3.2 REMD and Completeness of Sampling

A notable exclusion from these simulations is the tubulin protein itself, a factor that will undoubtedly influence the results as they have been explained here. The decision to exclude the bulk tubulin was made due to its large system size and the computational resources that would be required to explore the conformational space of an entire tubulin dimer. Therefore, simulations of only the CTTs become a problem of protein folding and adequate sampling is critical to the proper interpretation of the results. To increase sampling efficiency at different temperatures, we have used Replica Exchange MD (REMD), which does not provide information on the dynamics of the system. However, as our ultimate interest is in the overall conformations of the CTTs, the loss of dynamics data was an acceptable compromise in order to gain increased sampling. The completeness of sampling can be determined by calculating the normalized overlaps between two different parts of an MD trajectory. Such overlaps can indicate whether or not both parts are spanning over the conformational space equally and not diffusing to new parts [221–223]. Subspace overlap between two sets of n orthonormal vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ and $\mathbf{w}_1, \ldots, \mathbf{w}_n$ is defined as:

$$\text{overlap}\,(\mathbf{v}, \mathbf{w})\, n \sum_{i=1,j=1}^{n,n} (\mathbf{v} \cdot \mathbf{w})^2 \tag{8.4}$$

When the overlap has a value of 1, the sets $\mathbf{v}$ and $\mathbf{w}$ can be considered to span the same subspace. This measure, however, does not account for the magnitudes of the eigenvalues, meaning that differences between all eigenvectors contribute equally. Furthermore, when two or more eigenvalues are equal, the corresponding eigenvectors are random, causing a random variation in the subspace overlap number. An alternative method, suggested by Hess [223], is the normalized overlap, defined as:

$$d = \sqrt{\text{tr}\left(\left(\sqrt{C_1} - \sqrt{C_2}\right)\right)^2} \tag{8.5}$$

$$\text{normalized overlap}\,(C_1, C_2) = 1 - \frac{d}{\sqrt{\text{tr}\,(C_1) + \text{tr}\,(C_2)}} \tag{8.6}$$

where d is the difference between the covariant matrices, C_1 and C_2 , and tr is the trace. Here, if the overlap is 0, then the two sets are considered to be orthogonal, while an overlap of 1 indicates that the matrices are identical. Covariant analysis of the trajectories from each isotype at 311 K, chronologically divided into thirds, was performed using the same procedure used for the PCA. The subspace and normalized overlaps calculated between each of these thirds are reported in Table 8.3. The high overlap between the thirds indicates that each part of the simulation is sampling approximately the same conformational space and it is unlikely that there are unexplored regions missed earlier in the run. While not a guarantee of complete equilibrium sampling, we have concluded that the overlap using both of the above-mentioned methods is acceptable and that adequate sampling within all of these systems has been obtained

PCA Overlap at 311 K for N-Terminus Fit						
Isotype	Subspace Overlap			Normalized Overlap		
	1st vs. 2nd	1st vs. 3rd	2nd vs. 3rd	1st vs. 2nd	1st vs. 3rd	2nd vs. 3rd
I	0.93	0.95	0.95	0.86	0.84	0.86
II	0.90	0.87	0.92	0.84	0.77	0.78
III	0.94	0.91	0.95	0.82	0.78	0.86
IVa	0.92	0.84	0.91	0.77	0.62	0.75
IVb	0.93	0.89	0.89	0.78	0.76	0.83
V	0.92	0.89	0.89	0.75	0.78	0.76
VI	0.88	0.88	0.94	0.72	0.71	0.84
VII	0.96	0.91	0.98	0.70	0.56	0.76
VIII	0.92	0.86	0.94	0.79	0.76	0.85

Table 8.3: PCA overlap of CTT peptides. The subspace overlap consists of the first ten eigenvectors only. For both overlap calculations the ensemble was divided into thirds and all are compared to each other.

8.3.3 Clustering and Secondary Structure

Clustering each CTT peptide can provide a representation of probable folded conformations, however a bell shaped ρ_{sc} distribution about a representative structure demonstrates that there is actually no native folded conformation for most of the isotypes (Figure 8.2). The only exception was the βVII CTT, which exhibited a significant population of folded structures with a $\rho_{sc} < 0.5$ and a second population of unfolded structures with a $\rho_{sc} > 0.5$. These results recapitulate previous observations that the CTTs are extremely flexible and any structures within them are transient in nature and cannot be captured by the present clustering methodology. Therefore, we felt that a more appropriate analysis was to consider the ensemble average of secondary structures as calculated from the entire 10 ns of ensemble conformations at 311 K by STRIDE [224] (Figure 8.3). These results demonstrated that, while the ρ_{sc} distribution showed no native folded conformations, many of the CTTs contain regions that are either α- or 3-10-helical at least 40% of the time.

8.3.4 Motif Identification

Having established the presence of transient secondary structures within the ensembles, we performed pair-wise alignments of all the CTT sequences in order to identify potential sequence motifs with which to correlate our modeling results. A ρ_{sc} matrix was then calculated for all the conformations across each isotype at 311 K. Only those motifs having a high fraction of low ρ_{sc} were determined to be significantly similar (ρ_{sc} 0.3 or less) (Table 8.2). Through this analysis, we identified two motifs that showed both sequence and conformational similarities. To visually compare the motifs, the conformations of the aligned sub-sequences were clustered as described in Methods. Illustrated in Figure 8.4 are RMSD alignments of representative structures (those with the most neighbors) from each isotype containing the respective motif. The first of these motifs was identified as a probable Casein Kinase-2 (CK-2) binding motif at the N-terminal end of the peptide (Figure 8.4(a)). This motif was also independently identified using a Prosite search for motifs within each of the CTT sequences [225] (not shown). The second motif was determined to correspond to a previously identified MAP2 binding motif found within α-tubulin (Figure 8.4(b)) [226]. We should note that while a common conformation for each of the motifs across isotypes has been identified, these are not stable folds and depending on the isotype and residues included in the search, anywhere between 1 to 69% of the ensemble structures have a $\rho_{sc} < 0.3$ when compared to the motif conformation.

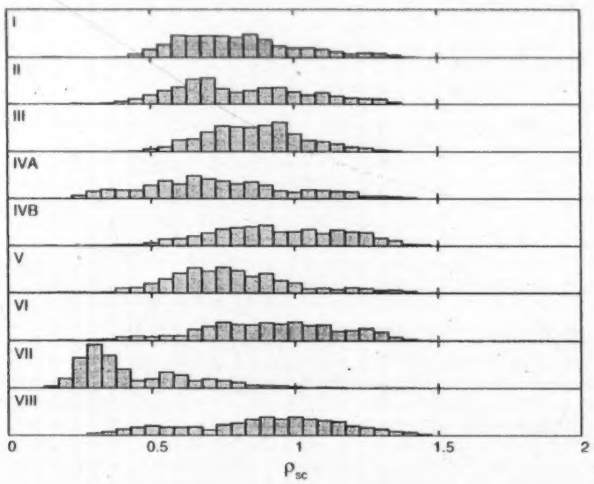


Figure 8.2: ρ_{sc} distributions of CTTs about representative structures. The large spread in the histograms of the ρ_{sc} distributions for each CTT indicates the lack of a single folded conformation. A distinct population of 67% conformers have a ρ_{sc} less than 0.5 for the βVII isotype, suggesting a distinct folded conformation for this CTT.

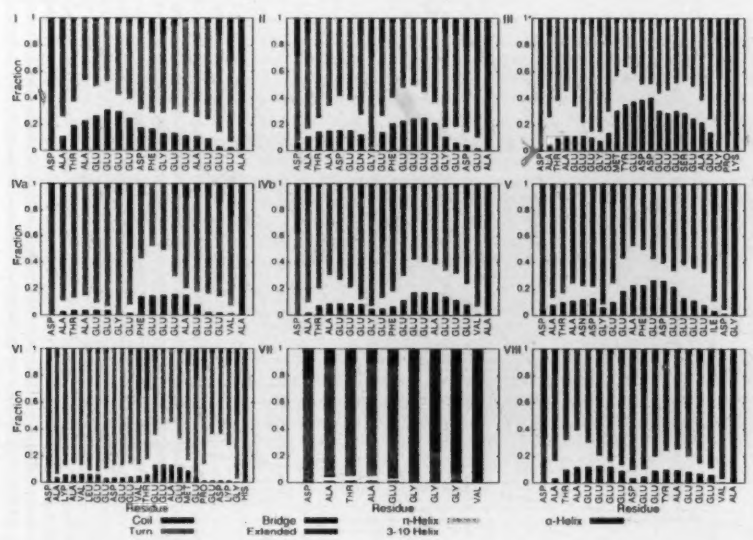
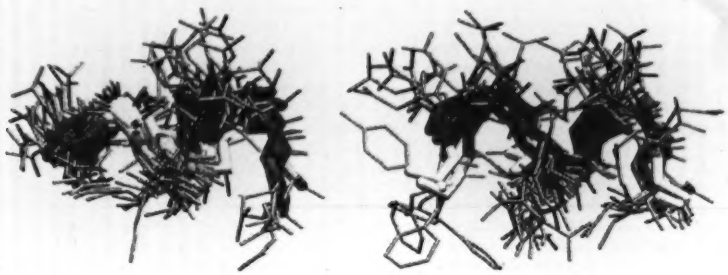


Figure 8.3: Time average of each type of secondary structure for each CTT. The total fractional secondary structure content was determined as an average over the entire 10 ns MD simulation at 311 K. The fractional time average of each type of secondary structure are stacked to sum up to one. While experimental studies of tubulin would suggest that the CTTs are unstructured, these results suggest that many of the CTTs contain a significant amount of transient helix.



(a) (b)

Figure 8.4: Sequence motif identification. High scoring sequence pairs from Table 8.2 were RMSD fit to illustrate the overall fold and position of side chains in the structural motifs. In both images, red residues are acidic, white residues are nonpolar, green residues are polar and all side chains are blue. Images created with VMD [11].

8.3.5 PCA

Not only is the secondary structure of each CTT significant, but their flexibility is also of critical importance when attempting to understand how they are able to conform when interacting with MAPs. Because of the complex interplay of many various factors, sophisticated statistical analysis of conformational ensembles, such as PCA, appears to be the most appropriate computational tool to characterize the flexibility. PCA was performed on the ensemble at 311 K, with the backbone atoms of the first three residues RMSD fitted to a reference structure (Figure 8.1). Resulting eigenvectors were ordered by descending eigenvalues, which represent the variance of the motion along the principal components. In Figure 8.5, the six largest root normalized eigenvalues are shown for each isotype. Except for βVII and βIVa, to a lesser extent, three components can be identified with prominent eigenvalues of comparable magnitude, which is significantly higher than the magnitude of other eigenvalues. The three components with the largest eigenvalues represent correlated motions of the peptide fragments with the most significant standard deviations of the motion along the corresponding orthogonal directions. Since the standard deviations shown in Figure 8.5 were normalized for the number of residues, they can be employed as a universal measure for comparison of conformational motions in different CTT peptides.

8.3.6 2D-Projections of the REMD conformation ensemble

A more detailed representation of the conformational motions in peptides is provided by projections of the ensemble of conformations onto the planes spanned by the most important principal components (Figure 8.6). Here, the spatial distributions of occupancies of the various conformational states are shown over the planes spanned by the first and second, and by the first and third principal components. A representative isotype, βV, illustrates that both distributions are smooth and without evidence of considerable clustering. Similar behavior exists in all isoforms considered except for βIVa and βVII, which show significant clustering of conformational occupancy, with maxima at (0.1,0) and (-0.2,0), respectively. The widths of the distributions shown in Figure 8.6 are not directly dependent on the length of the peptide, as the eigenvalues have been normalized by the number of residues. The reduced width of βVII, therefore, indicates its reduced mobility. The bin size of the histogram has been chosen in proportion to the width of the distribution so the maxima seen for βIVa and βVII are not

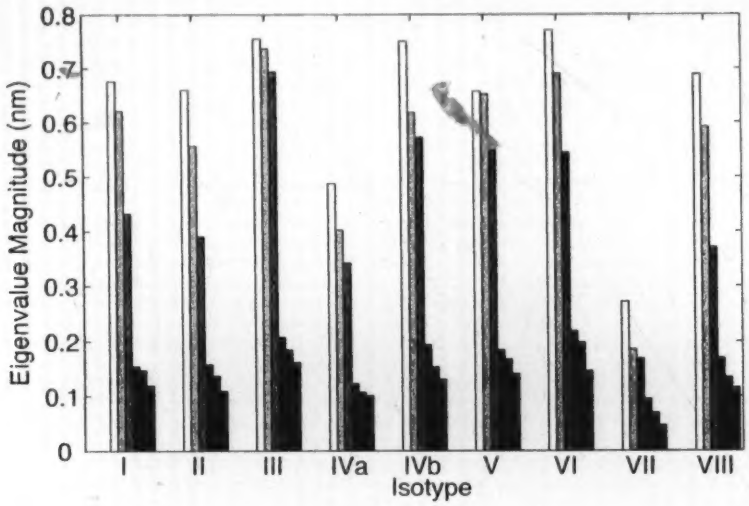


Figure 8.5: Root-mean-squared-normalized eigenvalues for each isotype. The magnitudes of the first six root normalized eigenvalues (white to black) are shown for each isotype and represent the standard deviation of the motion along the eigenvector normalized for the overall length of the peptide. The relative size of the root normalized eigenvalues indicates the relative flexibilities of the individual peptides.

artifacts of the sampling, but represent regions of increased occupancy. Since the occupancy of a conformational state is inversely proportional to the free energy of that state [227, 228], the lack of clustering in the other isotypes suggests that the global minima are broad or that there are generally no significant local minima. In contrast, βVII has a significant basin of attraction, indicative of a folded conformation. This observation is consistent with the preceding cluster analysis, which identified a well-defined folded state in the βVII isoform only. While less defined, the occupancy distribution for βIVa also exhibits a preferred folded conformation, demonstrating that the PCA methodology is more sensitive to transient structural motifs than the clustering analysis. One more noteworthy feature is that the 2D distributions over the 1st and 2nd components are largely similar to those over the 1^{st} and 3^{rd} components, suggesting a significant level of symmetry between principal components.

8.3.7 Relative CTT Flexibility

In addition to information regarding the motion of each CTT peptide, PCA also provides the ability to compare relative flexibilities. As we have only studied the nine human β-tubulin isotypes, it is not possible to comment on the flexibility of the CTT peptides in an absolute sense. However, there have been few studies discussing the flexibility of small peptides. While they use a different approach to calculating the flexibility, Ma et al. (2000) made a survey of 28 short peptides and determined that the native helical structures of the peptides were more flexible than random or disordered conformations, in agreement with our observations (Figure 8.7(b)) [229]. Unfortunately, their methodology, calculating the vibrational free energy, does not allow comparison of flexibility between peptides. Here, as a measure of the relative flexibility of each peptide the eigenvalues of principal components were compared. To

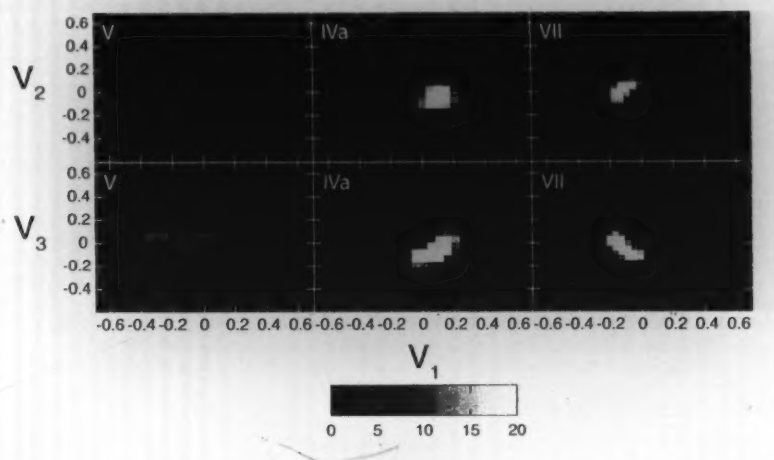


Figure 8.6: Projections of the ensemble of conformations onto the planes of the three most important principal components. The first and second principal components (upper row), and the first and third principal components (lower row) are plotted on the x- and y-axes respectively for βV, βIVa and βVII. The histograms represent the occupancies of the corresponding conformation states, with lighter colors indicating more frequently visited areas. The same coloring scheme is used for all isotypes and is capped at 20 counts. βV is typical of the remaining isotypes in the width of the distribution and lack of a significant maxima.

characterize the flexibility, we employed the value $\sigma_{12} = \sqrt{\sigma_1^2 + \sigma_2^2}$, where σ_1 and σ_2 are the normalized eigenvalues for the first and the second principal components, respectively. The magnitude of the PCA eigenvalue parameter σ_{12} can be viewed as the mean square width of the 2D REMD configuration ensembles shown in Figure 8.6, and thus σ_{12} can be directly interpreted as the flexibility of the peptide associated with its major correlated motions. Although the CTTs are characterized by three principal components (Figure 8.5), the considerable symmetry between the 2nd and 3rd component demonstrated in Figure 8.6 allows the use of only two components out of three to quantify the flexibility of CTTs. The correlations between the normalized distance (the distance between the N- and C-termini Cαs, divided by the number of residues), the time averaged helical content, and the average clustering parameter $\langle \rho_{sc} \rangle$ with the value σ_{12}, which we use as the major measure of flexibility, can be seen in Figure 8.7.

8.4 Discussion

As the CTTs properties are critically involved in MT regulation, it is essential to understand the conformational differences adopted by different tubulin isotypes. We suggest that the sequence variability within the CTTs may have arisen as a mechanism to conserve the overall tubulin structure in order to maintain proper MT assembly, while still providing a flexible, solvent exposed region of increased variability for interactions with proteins that affect MT function (see alignment in Table 8.1). This is both an attractive and reasonable hypothesis, as the role of intrinsically disordered proteins in protein interactions has been implicated as a mechanism to enhance specificity [231]. The following discussion will address three main points with regards to the CTTs and their interactions with MAPs: first, the relative flexibilities of each of the CTT peptides, second, the presence of transient structure within these peptides,

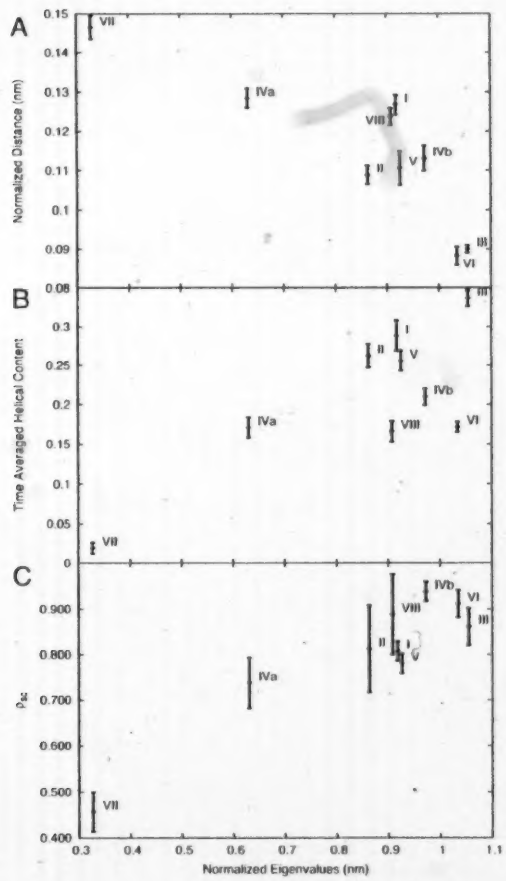


Figure 8.7: Peptides flexibility indicators. The end-to-end distance normalized by the number of residues (a), time average helicity (b) and $\langle \rho_{sc} \rangle$ (c) of each CTT peptide at 311 K plotted as functions of σ_{12}. The plots demonstrate the correlation between each of the above values and the major flexibility indicator, σ_{12}. In (b) βVI can be excluded from the trend due to the presence of a Pro, which disrupts the secondary structure but contributes to the flexibility of the peptide. The error in all three plots is the standard error calculated using box averaging [230].

and finally how this may influence protein interactions. We will discuss this possibility further when examining CTT flexibility in the context of MAP interactions, in particular interactions with the CK-2 and MAP2 binding motifs, that we have identified in this work.

8.4.1 CTT Flexibility and Secondary Structure

Accurate flexibility and secondary structure measurements are critical when the goal is to propose mechanisms for MAP interactions with the large number of possible CTT conformations on the MT surface. The overall distribution of amino acids within the CTT will obviously have an effect on their flexibility and result in secondary structure. For example, the overall helical propensity of βII, βIII, βIVa, βIVb and βV is interrupted by the presence of Gly (Figure 8.3). Additionally, regions with uninterrupted stretches of Glu also tend to produce regions with greater helical propensity, providing an argument for the prevalence of this residue within the CTT sequences. This observation can be compared with the results of Roe et al. (2007) for simulations performed on deca-alanine [77]. While shorter than all but the βVII peptide, the overall α- and 3-10 helical content within deca-alanine was similar to the helical content observed here. However, through the analysis of the distribution of secondary structure along the CTT sequences, we note that our results differ from that of Roe et al. (2007), in that the helical content of the CTTs drops near the middle of the sequence for most isotypes while for deca-alanine the central residues have a maximal amount of helix.

Three characteristics emerge which are representative of the behaviors of CTT; the end-to-end distance, normalized by the number of residues, the PCA-based eigenvalues, and the averaged helical content (Figure 8.7, panels (a) and (b)). To test the applicability of PCA as a measure of relative flexibility, we have compared PCA-based flexibilities with the average parameter of structural similarity $\langle \rho_{sc} \rangle$, which we have computed from the distributions in Figure 8.2 (Figure 8.7(c)). Although less detailed than PCA, the distributions of conformations over the parameter ρ_{sc} still provide an alternative statistical characterization of the CTTs configurations, and thus they can be expected to correlate with the PCA-based results. The correlation between the PCA-based flexibilities and $\langle \rho_{sc} \rangle$ is more pronounced than any other correlation that we have investigated. Most of the CTTs show a clear proportionality between σ_{12} and $\langle \rho_{sc} \rangle$, and the variability of $\langle \rho_{sc} \rangle$ that corresponds to similar σ_{12} is less than 15%. This similarity of results from two fundamentally different statistical tools demonstrates that these statistical methodologies are the best suited to characterize CTTs flexibility. Interestingly, rather than the expected exponential decay, the magnitudes of the first three eigenvalues are of similar magnitude and significantly larger than the remaining eigenvalues, indicating an isotropy of occupancies in the 3D space (Figure 8.5).

More simply, the CTTs are flexible enough for the ensemble of their configurations to be spherically symmetric with respect to an immobilized base of amino acids (Figure 8.1). Since these three eigenvalues are significantly larger than all the others, 3D symmetry is likely reached through a few highly flexible bonds within the peptides that allow rotations of large angles, thereby generating nearly spherically symmetric distributions of occupancies. In Figure 8.6 this is illustrated by a comparison of the histogram for βV, which is representative of a significant level of symmetry typical for most isotypes, with less symmetric ones for βIVa and βVII. As the magnitudes of the eigenvalues shift to an exponential decay, spherical symmetry is lost and correlated motions begin to appear. At the same time, an individual CTT itself may preserve relatively stable motifs contained within some secondary structure. In terms of the PCA results presented in Figure 8.5, these motifs are represented by the fourth, fifth, and higher eigenvalues. The fact that these components, although significantly less pronounced than the first three, still have considerable nonzero eigenvalues, indicates that the conformational motifs within the CTTs are of transient nature and subject to variability over the trajectory. This is consistent

with our observation of a significant amount of transient helicity (Figure 8.3) indicated by a substantial amount of 3-10 helix.

Considering the PCA eigenvalue parameter σ_{12} as the measure of flexibility, it is evident that the βIII and βVI CTTs are the most flexible of all CTTs. These are the longest two sequences, which may account for some of their flexibility, but interestingly they also have the shortest end-to-end distance per residue. The cause of these shared structural properties differs in the two cases, while the βIII CTT has among the highest degree of secondary structure, the βVI CTT has among the lowest. The βVI CTT differs from all the other fragments as it contains a Pro near the C-terminus of the peptide. This Pro breaks the helical structure of the fragment and, at the same time, does not facilitate an extended conformation. This accounts for the lack of secondary structure but is in contrast to βIII, which predominantly lies in the α-helical region of the ϕ/ψ distribution. When accounting for the lone Pro and omitting the data for βVI, the correlation between helicity and flexibility becomes greater. The βII, βIVb and βV CTTs also show the same spring-like flexibility of βIII, exhibiting a similar helical content and moderate end-to-end distance per residue. These CTTs also contain Gly at the same position that breaks the helical content of the fragment. In contrast, the βIVa and βVII CTTs displayed the least amount of flexibility. In the case of βIVa, one reason for this reduction in flexibility may be the early occurrence of Gly in its sequence, which stiffens this region of the peptide by decreasing the helicity. This decrease in helicity in the first seven residues, compared to βIVb, can be seen in Figure 8.3. βIVa also contained the most stable DA[TK]AEE motif (data not shown), which resulted in the largest normalized distance of all the common isotypes.

The observation that peptides containing a significant amount of transient helical conformations are more flexible than an extended one may seem counter intuitive. The correlation between the normalized distance and flexibility in Figure 8.7(a) is therefore most easily understood in terms of entropy. Take the following example: the freely jointed, or ideal chain, polymer model states that compact forms of the polymer contain more accessible states (i.e. have greater entropy) than do extended states [232]. Thus, an entropic force drives the polymer or polypeptide into compact conformations. A greater number of accessible states suggests that the chain is more flexible and that extended states are more rigid. While the ideal chain has no internal energy and is free to collapse, a peptide fragment does contain internal energy that can stiffen the peptide. The source of the force that causes the peptides to be in a rigid, extended state is the peptide backbone. Individual residues that induce a compact conformation through helical propensity or residues that induce a turn (such as proline) provide flexibility to the structure of the peptide while reducing the normalized distance. However, there is not a complete one-to-one correspondence between the flexibility and the normalized distance. For example, βI and βVIII have a normalized distance comparable to βIVa but a σ_{12} closer to the more compact βV. While the normalized distance of the peptide is a strong indicator of flexibility the details of the compact structure, such as the degree of helicity, also contribute. We are not the first to observe this phenomenon in peptides, Ma et al. (2000) calculated the vibrational free energy of 28 short peptides in native helical and random extended states and determined that the native helical structures of the peptides were more flexible in all cases [229].

It is becoming clear that each β-tubulin isotype has a unique pattern of expression ranging from specific for βII, βIII, βIVa and βVI, to constitutive for βI and βIV [233, 234]. The βI isotype is the most commonly expressed in humans and, as such, is also the most common isotype found in cancer cells [235]. Our observations also suggest that the CTT of this isotype shows "typical" behavior in our analysis. It is difficult to correlate our results to biological function, however a clear pattern illustrating βIII as a statistical outlier has emerged. The βIII CTT was one of the most flexible peptides and contained a large amount of transient

secondary structure. This is interesting, as βIII-tubulin has been observed at increased levels in human tumors and implicated in the development of drug resistance to standard chemotherapy treatments [236–238]. In addition to the altered expression of βIII, the alteration of tubulin regulatory proteins, such as MAP4 has also been implicated in changes to MT dynamics and the development of drug resistance [239]. Because several MAPs have now been shown to interact directly with the CTTs [60, 207–210], the differences that we have observed here may have a significant impact on their affinity to the surface of a MT.

8.4.2 Motifs and MAP interactions

Using the conformational similarities for each CTT peptide, in combination with sequence comparisons, we have identified two distinct transient structural motifs (Table 8.2). Several similar regions were observed as significant conformations throughout the trajectory files indicating that they may be common motifs that could potentially be recognized by MAPs. The first motif was identified as a putative casein kinase II phosphorylation site (CK-2) for which the consensus motif has been identified as [ST]XX[DE] [225]. All of the CTTs, with the exception of βVI and βVII, were shown to contain this motif within the conserved DA[TK]A sequence. The second motif identified was the MAP2 binding site, for which the consensus sequence has previously been characterized as EEAEEE [226]. Only the CTT of βI, βIVa and βIVb, strictly contain this motif, however most of the other CTTs also contain a similar motif (Table 8.2). The conformations identified for βI, βIVa, βIVb, and βVI most commonly form a transient helix that is between 4 and 5 residues in length. In βVIII, there seems to be an increasing amount of unstructured loop associated with this region, which is most likely a result of the displacement of the amino end due to the presence of a bulky Tyr residue. Both the CK-2 and MAP2 motifs display a significantly higher degree of helical propensity when compared to the entire ensemble of CTT conformations (Figure 8.3). Interestingly, with the exception of βVI and βVII, the increased helical propensity of the proposed CK-2 domain suggests that the H12 helix may extend several additional residues farther than those that have already been observed in the crystal structures of tubulin.

Finally, the motifs identified here are not stable structures but rather commonly occurring, transient conformations. For example, in the "stiffest" isotype, βIVa, the CK-2 DATA motif is occupied 69% of the time while the MAP2 motif EEEV is occupied 17% of the time (Table 8.2). The two motifs are simultaneously folded in the ensemble of CTT trajectories only 12% of the time. As the sequences contained within the two motifs are extended to include DATAEE and AEEEVA, respectively, this drops to 2% (Table 8.2). Rather, the shared similarity of the folded states indicates that these conformations are individually inducible, particularly when binding to their respective substrates. As we are dealing with a subset of all CTTs, we were unable to determine accurately if any additional structural or sequence motifs exist in the tails. A more accurate determination of CTT motifs would therefore require a larger survey of CTTs across all identified α- and β-tubulin isotypes (see Table 8.1).

8.5 Conclusions

The role of the unstructured tubulin CTTs, particularly concerning MT stability and interactions with MAPs, is a question that has yet to be conclusively answered. Here we have shown, using a combination of REMD and PCA, that while the experimental structures of tubulin suggest the CTTs are unstructured, many contain a significant amount of transient conformations that are linked by a few highly rotatable bonds. We believe that the role of the transient secondary structure, coupled with increased CTT flexibility may be twofold. First,

global flexibility, coupled with weak helical tendencies could provide a mechanism that allows the CTTs to search conformational space and therefore easily bind to other proteins. This could provide a convenient mechanism for MAPs to dock to the MT surface. Second, increased flexibility could also enhance binding affinities to MAPs as a result of their ability to conform to a specific binding site. The presence of well-defined structural motifs within the CTTs may therefore play an important role in binding specificity, where the presence of smaller domains provides a common structural signal for binding, while the global flexibility of the CTT is maintained at the same time.

The role of β-tubulin CTTs in MT function is appealing, since subtle differences in a small stretch of amino acids could have profound consequences on MAP interactions. Unfortunately, it is extremely challenging at the present time to elicit tight correlations between the results presented here and their biological consequences, as there is very little experimental data regarding CTTs explicitly in the context of tubulin isotypes. As interest in the properties of tubulin isotypes and their potential role in targeted chemotherapy treatments increases, we anticipate a significant body of data soon to become available. As a result, a clear understanding of CTT structure and function may provide the means by which we can begin to rationally design and develop novel drugs or peptide mimetics that target CTT binding sites, providing better selectivity for chemotherapies.

Chapter 9

Molecular Theory of Solvation

Accurate modelling of solvation effects is essential to to correctly describe the physical properties of biomolecules. Chapter 4 detailed important properties of water, how they influence proteins and common models. Models are generally divided into two groups, explicit and implicit, balancing accuracy against computational efficiency.

Implicit models neglect the molecular basis of water's properties and opt for a macroscopic continuum treatment. Due to the lack of a molecular basis, the primary way these models differentiate between solvents like methanol, octanol and water is through their respective dielectric constants. In fact, these models do not account for the full solvation free energy, calculating only the free energy of solvent polarization. Despite their shortcomings, these models do provide a quantitative account of solvent polarization, remove the solvent degrees of freedom, and often, can be calculated quickly and efficiently.

Explicit models offer the most accurate account of solvation at the molecular level. These models generally have problems with reproducing accurate dielectric constants, diffusion constants, water structure and temperature dependence. Even with these short comings, they have been demonstrated capable of reproducing solvation effects in adequate detail for the simulation of biomolecules. In fact, the main draw back of these models is often considered to be the computational cost rather than the quantitative accuracy. Often a three-point model will be used to save time over a more accurate four or five-point model. This is not surprising as water typically accounts for 95% of the atoms in the system. Furthermore, the increased degrees of freedom associated with the solvent can be detrimental to sampling techniques, such as replica exchange, that scale with respect to the number of degrees of freedom.

Ideally, we would like a method that combines the computational efficiency of implicit models with the accuracy of explicit models. A possible solution, explored in this Chapter and Chapter 10, is the 3D-reference interaction site model (3D-RISM) of solvation. It is uses a molecular description of the solvent (i.e. an explicit model) to calculate the ensemble properties of the solvent, thereby contracting the degrees of freedom, just as with implicit models. As will be demonstrated, 3D-RISM compares favourably in reproducing properties of explicit solvent models. The major issue facing 3D-RISM coupled with MD is that of speed.

This Chapter and Chapter 10 together detail the implementation, optimization and characterization of 3D-RISM coupled to the Amber molecular dynamics (MD) package. Section 9.1 describes the background theory of 3D-RISM. Section 9.2 gives details of the implementation and various approaches to optimizing 3D-RISM in Amber. Section 9.3 and Chapter 10 focus on the characterization of both the quality of the 3D-RISM output and the performance.

9.1 Theoretical Background

In addition to explicit and implicit solvent models, a third method to be considered is the reference interaction site model of molecular solvation. This model starts with a microscopic approach, as opposed to the macroscopic one used by GB and PB, and is better able to handle surface molecular specificity effects at the solvent-solute boundary.

Though RISM is now nearly 40 years old it is still much younger than PB, GB or explicit solvent models. It has gone through three main stages of development. The first was Chandler and Andersen's original RISM theory, published in 1972 [240]. This was followed a decade latter with the development of extended RISM (XRISM) by Hirata, Rossky and Pettitt [241–243]. Until this point, RISM was a site-site 1D theory; it averaged the orientations of both the solute and solvent, making its application to solutes of complex geometry difficult. In the 1990s, Beglov and Roux [1, 2] and Kovalenko and Hirata [3, 4] further extended RISM to 3D, creating 3D-RISM. Because only the solvent is orientationally average, 3D-RISM can be applied to biomolecules and, with some effort, to MD (see [244] and Chapter 10), Monte Carlo [245–247] and quantum chemistry calculations [248].

9.1.1 Ornstein-Zernike Equation

The RISM methodology is based on the Ornstein-Zernike (OZ) integral equations for a complex fluid [4, 232, 240, 249, 250]. In general terms, this equation may be written as

$$h(r_{12},\Omega_1,\Omega_2) = c(r_{12},\Omega_1,\Omega_2) + \rho \int d\mathbf{r}_3\, d\Omega_3\, c(r_{13},\Omega_1,\Omega_3)\, h(r_{32},\Omega_3,\Omega_2), \tag{9.1}$$

where r_{12} is the separation between particles 1 and 2 while Ω_1 and Ω_2 are their orientations relative to the vector $\mathbf{r}_{12}$. The two functions in this relation are h, the total correlation function, and c, the direct correlation function. The total correlation function is defined as

$$h_{ab}(\mathbf{r}) \equiv g_{ab}(\mathbf{r}) - 1, \tag{9.2}$$

where g_{ab} is the pair-distribution function, which gives the conditional density distribution of b about a. Orientationally averaging over the sites, in 1D, this is the familiar site-site radial distribution function. The direct correlation function is, in the low density limit, the correlation between two particles separated by distance r_{12} with orientations Ω_1 and Ω_2 [232]. That is, $h = c$ when there are only two particles in our system. As the density of our system increases, the effects to 3-body interactions and higher are included via the integral of Equation (9.1).

For real mixtures, it is often convenient to speak in terms of a solvent, V, of high concentration and a solute, U, of low concentration. We can rewrite Equation (9.1) relation as a set

of three equations:

$$h^{VV}(r_{12},\Omega_1,\Omega_2) = c^{VV}(r_{12},\Omega_1,\Omega_2) + \rho^V \int d\mathbf{r}_3\, d\Omega_3\, c^{VV}(r_{13},\Omega_1,\Omega_3)\, h^{VV}(r_{32},\Omega_3,\Omega_2), + \rho^U \int d\mathbf{r}_3\, d\Omega_3\, c^{VU}(r_{13},\Omega_1,\Omega_3)\, h^{UV}(r_{32},\Omega_3,\Omega_2), \quad (9.3)$$

$$h^{UV}(r_{12},\Omega_1,\Omega_2) = c^{UV}(r_{12},\Omega_1,\Omega_2) + \rho^V \int d\mathbf{r}_3\, d\Omega_3\, c^{UV}(r_{13},\Omega_1,\Omega_3)\, h^{VV}(r_{32},\Omega_3,\Omega_2) + \rho^U \int d\mathbf{r}_3\, d\Omega_3\, c^{UU}(r_{13},\Omega_1,\Omega_3)\, h^{UV}(r_{32},\Omega_3,\Omega_2) \quad (9.4)$$

$$h^{UU}(r_{12},\Omega_1,\Omega_2) = c^{UU}(r_{12},\Omega_1,\Omega_2) + \rho^V \int d\mathbf{r}_3\, d\Omega_3\, c^{UV}(r_{13},\Omega_1,\Omega_3)\, h^{VU}(r_{32},\Omega_3,\Omega_2), + \rho^U \int d\mathbf{r}_3\, d\Omega_3\, c^{UU}(r_{13},\Omega_1,\Omega_3)\, h^{UU}(r_{32},\Omega_3,\Omega_2). \quad (9.5)$$

The most interesting case of solvation is infinite dilution of the solute, i.e. $p^U \to 0$. The third term on the R.H.S. of these three equations vanishes and we are left with

$$h^{VV}(r_{12},\Omega_1,\Omega_2) = c^{VV}(r_{12},\Omega_1,\Omega_2) + \rho^V \int d\mathbf{r}_3\, d\Omega_3\, c^{VV}(r_{13},\Omega_1,\Omega_3)\, h^{VV}(r_{32},\Omega_3,\Omega_2), \quad (9.6)$$

$$h^{UV}(r_{12},\Omega_1,\Omega_2) = c^{UV}(r_{12},\Omega_1,\Omega_2) + \rho^V \int d\mathbf{r}_3\, d\Omega_3\, c^{UV}(r_{13},\Omega_1,\Omega_3)\, h^{VV}(r_{32},\Omega_3,\Omega_2), \quad (9.7)$$

$$h^{UU}(r_{12},\Omega_1,\Omega_2) = c^{UU}(r_{12},\Omega_1,\Omega_2) + \rho^V \int d\mathbf{r}_3\, d\Omega_3\, c^{UV}(r_{13},\Omega_1,\Omega_3)\, h^{VU}(r_{32},\Omega_3,\Omega_2), \quad (9.8)$$

However, these equations can not yet be solved. c is defined by these equations in terms of h, which is defined by g, which is unknown. Thus, to solve this system of equations we need to introduce one more, a closure relation.

9.1.1.1 Closure

Equations (9.6) to (9.8) are single equations with two unknowns each. To obtain a solution it is necessary to have a second equation that relates h and c or uniquely defines one of these functions. The general closure relation is [249]

$$c(r_{12},\Omega_1,\Omega_2) = \exp\left[-\beta u(r_{12},\Omega_1,\Omega_2) + t(r_{12},\Omega_1,\Omega_2) + b(r_{12},\Omega_1,\Omega_2)\right] - 1 - t(r_{12},\Omega_1,\Omega_2) \quad (9.9)$$

where

$$t(r_{12},\Omega_1,\Omega_2) = h(r_{12},\Omega_1,\Omega_2) - c(r_{12},\Omega_1,\Omega_2) \quad (9.10)$$

u is the potential energy function for the two particles and b, a functional of h and c, is known as the bridge function. In principle, it should now be possible to solve our two equations. The bridge function, however, contains multiple integrals that are readily solved only in special circumstances. An approximate closure relation must be used in practice and, ideally, should be based on some physical grounds.

Several closure relations have been developed. Amongst the most notable are the hypernetted chain equation (HNC) [249]

$$c(r_{12},\Omega_1,\Omega_2) = \exp\left[-\beta u(r_{12},\Omega_1,\Omega_2) + t(r_{12},\Omega_1,\Omega_2)\right] - 1 - t(r_{12},\Omega_1,\Omega_2) \quad (9.11)$$

where we have set $b = 0$, Percus-Yevick (PY) [232, 249]

$$c(r_{12}, \Omega_1, \Omega_2) = \exp\left[-\beta u(r_{12}, \Omega_1, \Omega_2)\right]\left[1 + t(r_{12}, \Omega_1, \Omega_2)\right] - 1 - t(r_{12}, \Omega_1, \Omega_2) \tag{9.12}$$

which is a partial linearization of HNC, and the mean spherical approximation [249]

$$h(r) = -1 \text{ for } r \leq \sigma \tag{9.13a}$$

$$c(r) = -\beta w(r) \text{ for } r > \sigma \tag{9.13b}$$

where σ is the particle radius and $w(r)$ is the attractive (or repulsive) tail of the potential

$$u(r) = \begin{cases} \infty & \text{for } r \leq \sigma \\ w(r) & \text{for } r > \sigma \end{cases}$$

All three of these closures have deficiencies. HNC works well in many situations but has difficulties when the size ratios of particles in the system are highly varied. PY generally has difficulty with charged systems, manifesting in a negative radial distribution function at short separations. MSA also has similar difficulties.

A means to overcome some of these issues is the Kovalenko-Hirata (KH) closure, a combination of HNC and MSA [4]

$$g(r_{12}, \Omega_1, \Omega_2) = \begin{cases} \exp\Big(\mathcal{X}(r_{12}, \Omega_1, \Omega_2)\Big) & \text{for } \mathcal{X}(r_{12}, \Omega_1, \Omega_2) \leq 0 \\ 1 + \mathcal{X}(r_{12}, \Omega_1, \Omega_2) & \text{for } \mathcal{X}(r_{12}, \Omega_1, \Omega_2) > 0 \end{cases} \tag{9.14}$$

where

$$\mathcal{X}(r_{12}, \Omega_1, \Omega_2) = -\beta u(r_{12}, \Omega_1, \Omega_2) + h(r_{12}, \Omega_1, \Omega_2) - c(r_{12}, \Omega_1, \Omega_2)$$

Equation (9.14) and its first derivative are continuous. This formulation has the advantage of selecting MSA for regions of high density ($g > 1$) and HNC for regions of low density ($g < 1$) at intermediate to high densities. The KH closure holds the features of both a proper description of high association peaks and long-range tails, peculiar to MSA, as well as a correct representation of the repulsive core by HNC.

9.1.1.2 Solvation Free Energy

The solvation free energy at infinite dilution is the quantity that determines the stability of a solute in a solvent [251]. In a classical system, such as ours, the solvation free energy is equivalent to the excess chemical potential of the solute (excess from the ideal gas)[1]. From the solvation free energy all other excess properties may be calculated. This is defined as the free energy change associated with coupling a solute to a solvent. The coupling is expressed as

$$E(\lambda) = E^{VV} + E^{UV}(\lambda), \tag{9.15}$$

where E is the interaction potential energy of the system and λ is the Kirkwood coupling factor and takes on values from 0 (no interaction) to 1 (full solvation).

The free energy change associated with this is given as

$$\begin{aligned} \Delta\mu &= F(\lambda = 1) - F(\lambda = 0) \\ &= -k_B T \left(\ln Z(\lambda = 1) - \ln Z(\lambda = 0)\right), \end{aligned} \tag{9.16}$$

[1] Quantum systems also have to account for the polarization of the electron distribution in the solvent.

where Z is the configuration integral of the system. Z can be eliminated and Equation (9.16) can be expressed as

$$\Delta\mu = \sum_i \sum_\alpha \rho^V \int_0^1 d\lambda \int_0^\infty d\mathbf{r} u_{i\alpha}(\mathbf{r}) g_{i\alpha}(\mathbf{r}, \lambda), \tag{9.17}$$

where i is the i^{th} solute particle. This expression in quite general but practically suffers due to the integral over λ. Fortunately, solutions for this problem have been developed for specific closures for both 1D- and 3D-RISM.

9.1.2 1D-RISM

As Equations (9.6) to (9.8) depend on the orientations of both the solvent and solute, as well as the their separation, they are 6D equations and a challenge to work with. By site-site averaging over the orientations of the solute and solvent, these equations are reduced to a 1D dependence on distance alone. This allows an interaction-site model (ISM) to be used, as is commonly used in molecular simulation, and only the scalar distance between sites need be considered.

However, even simple molecules, like water, are not spherical and have an internal structure we would like to account for. This is accomplished with the main assumption of RISM, that the direct correlation function is decomposable into the sum of site-site direct correlation functions

$$c(r) = \sum_{\alpha\gamma} c_{\alpha\gamma}(|\mathbf{r}_{\alpha 1} - \mathbf{r}_{\gamma 2}|)) \tag{9.18}$$

where α and γ are interaction sites. This leads to the following relation in reciprocal space

$$c(k) = \sum_{\alpha\gamma} \tilde{c}_{\alpha\gamma}(k) \omega_{\alpha\gamma}(k) \tag{9.19}$$

where ω is the $\omega_{\alpha\gamma}(k) = \sin(kr_{\alpha\gamma})/(kr_{\alpha\gamma})$ is the Fourier transform of the intramolecular pair correlation function and $\tilde{c}$ is the orientationally averaged c.

This leads to the 1D-RISM analog of Equation (9.1)

$$\begin{aligned} h_{\alpha\gamma}(r) &= \frac{1}{(2\pi)^3} \int \left[\boldsymbol{\omega}\mathbf{c}\left[\mathbf{1} - \boldsymbol{\rho}\boldsymbol{\omega}\mathbf{c}\right]^{-1}\boldsymbol{\omega}\right]_{\alpha\gamma} e^{i\mathbf{k}\cdot\mathbf{r}} d\mathbf{k} \\ &= \sum_0^\infty \boldsymbol{\omega}(k)\mathbf{c}(k)\boldsymbol{\omega}(k)\left[\boldsymbol{\rho}\mathbf{c}(k)\boldsymbol{\omega}(k)\right]^n \end{aligned} \tag{9.20}$$

in matrix notation this is

$$\mathbf{h} = \mathbf{w} * \mathbf{c} * \mathbf{w} + \boldsymbol{\rho}\mathbf{w} * \mathbf{c} * \mathbf{h}, \tag{9.21}$$

where $\mathbf{w} = \boldsymbol{\omega}/\rho$ and in reciprocal space

$$h_{\alpha\gamma}(k) = \omega_{\alpha\mu}(k)\,\tilde{c}_{\mu\nu}(k)\,\omega_{\nu\gamma}(k) + \rho\omega_{\alpha\mu}(k)\,\tilde{c}_{\mu\nu}(k)\,h_{\nu\gamma}(k) \tag{9.22}$$

As with the OZ equation, this can be decomposed into interactions between solvent and solute [243]

$$\mathbf{h}^{VV} = \mathbf{w}^V * \mathbf{c}^{VV} * \mathbf{w}^V + \boldsymbol{\rho}^V \mathbf{w}^V * \mathbf{c}^{VV} * \mathbf{h}^{VV} + \boldsymbol{\rho}^U \mathbf{w}^V * \mathbf{c}^{VU} * \mathbf{h}^{UV} \tag{9.23}$$

$$\mathbf{h}^{UV} = \mathbf{w}^U * \mathbf{c}^{UV} * \mathbf{w}^V + \boldsymbol{\rho}^V \mathbf{w}^U * \mathbf{c}^{UV} * \mathbf{h}^{VV} + \boldsymbol{\rho}^U \mathbf{w}^U * \mathbf{c}^{UU} * \mathbf{h}^{UV} \tag{9.24}$$

$$\mathbf{h}^{UV} = \mathbf{w}^U * \mathbf{c}^{UU} * \mathbf{w}^U + \boldsymbol{\rho}^V \mathbf{w}^U * \mathbf{c}^{UV} * \mathbf{h}^{V\!U} + \boldsymbol{\rho}^U \mathbf{w}^U * \mathbf{c}^{UU} * \mathbf{h}^{UU} \tag{9.25}$$

Once again, these equations reduce for the case of infinite dilution

$$\mathbf{h}^{VV} = \mathbf{w}^{V} * \mathbf{c}^{VV} * \mathbf{w}^{V} + \boldsymbol{\rho}^{V}\mathbf{w}^{V} * \mathbf{c}^{VV} * \mathbf{h}^{VV} \tag{9.26}$$

$$\mathbf{h}^{UV} = \mathbf{w}^{U} * \mathbf{c}^{UV} * \mathbf{w}^{V} + \boldsymbol{\rho}^{V}\mathbf{w}^{U} * \mathbf{c}^{UV} * \mathbf{h}^{VV} \tag{9.27}$$

$$\mathbf{h}^{UV} = \mathbf{w}^{U} * \mathbf{c}^{UU} * \mathbf{w}^{U} + \boldsymbol{\rho}^{V}\mathbf{w}^{U} * \mathbf{c}^{UV} * \mathbf{h}^{VU} \tag{9.28}$$

With these equations and a closure relation, it is possible to calculate the thermodynamic properties of a given potential, such as the SPC/E model of water.

9.1.2.1 Solvation Free Energy

Equation (9.17) provides the solvation free energy for a solute at infinite dilution. However, in this form numeric integration of λ is required to obtain a value, which is not practical in general. Fortunately, there are analytical solutions for this integral for some specific closures.

Notably, Singer and Chandler [252] solved this for the HNC closure by showing there exists an exact differential of the solvation free energy (Equation (9.16)), obtaining the relation

$$\Delta\mu = 4\pi\rho^{V}k_{B}T\sum_{i}\sum_{\alpha}\int\left[\frac{1}{2}\left(h_{i\alpha}^{UV}(r)\right)^{2} - c_{i\alpha}^{UV}(r) - \frac{1}{2}h_{i\alpha}^{UV}(r)c_{i\alpha}^{UV}(r)\right]r^{2}dr. \tag{9.29}$$

While there does not exist an exact differential of the free energy for the PY or MSA closures, there does exist an exact differential for the KH closure [4, 250]:

$$\Delta\mu = 4\pi\rho^{V}k_{B}T\sum_{i}\sum_{\alpha}\int\left[\frac{1}{2}\left(h_{i\alpha}^{UV}(r)\right)^{2}\Theta\left(-h_{i\alpha}^{UV}(r)\right) - c_{i\alpha}^{UV}(r) - \frac{1}{2}h_{i\alpha}^{UV}(r)c_{i\alpha}^{UV}(r)\right]r^{2}dr. \tag{9.30}$$

Note that the Heaviside function, Θ, effectively employs the h^2 term only in areas of density depletion. The fact that there is an analytic expression for the solvation energy for this closure is of considerable importance as it allows the rapid calculation of thermodynamic properties and, as we shall see, solvation forces.

9.1.3 3D-RISM

While 1D-RISM works well for small, spherical molecules, problems arise for large complex molecules. Information lost through site-site orientational averaging prevents useful application of 1D-RISM to molecules like proteins. For such molecules it is essential to maintain a detailed 3D representation, even if it is not necessary for the solvent.

3D-RISM [1, 4, 250], like 1D-RISM, performs orientational averaging of the OZ equation. However, this orientational averaging only involves the solvent and not the solute. That is, only Equation (9.6) is orientationally averaged, becoming Equation (9.26). However, since h^{VV} is used by Equations (9.7) and (9.8), these equations are reduced from 6D to 3D and, rather than calculating the radial distribution function of the solvent about the solute, the 3D solvent pair distribution function is calculated. This gives the 3D-RISM analog of Equation (9.7)

$$h_{\gamma}^{\mathrm{UV}}(\mathbf{R}) = \sum_{\alpha}\int d\mathbf{R}' c_{\alpha}^{\mathrm{UV}}(\mathbf{R}-\mathbf{R}')\,\chi_{\alpha\gamma}^{\mathrm{VV}}(R'). \tag{9.31}$$

$c_{\alpha}^{\mathrm{UV}}(\mathbf{R})$ has the asymptotics of the solute-solvent site interaction potential:

$$c_{\alpha}^{\mathrm{UV}}(\mathbf{R}) \propto -u_{\alpha}^{\mathrm{UV}}(\mathbf{R})/(k_{B}T). \tag{9.32}$$

$\chi^{VV}_{\alpha\gamma}(R)$ is the site-site susceptibility of the solvent, given by

$$\chi^{VV}_{\alpha\gamma}(R) = \omega^{VV}_{\alpha\gamma}(R) + \rho_\alpha(h^{VV}_{\alpha\gamma}(R)), \tag{9.33}$$

where $h^{VV}_{\alpha\gamma}(R)$ obtained from Equation (9.26) and $\omega^{VV}_{\alpha\gamma}(R)$ is the same as used in the 1D-RISM calculation. As Equation (9.31) involves a convolution, it is convenient to express it in reciprocal space

$$h^{UV}_\gamma(\mathbf{k}) = \tilde{c}^{UV}_\alpha(\mathbf{k})\,\omega_{\alpha\gamma}(k) + \rho^V\,\tilde{c}^{UV}_\alpha(\mathbf{k})\,h^{VV}_{\alpha\gamma}(k). \tag{9.34}$$

In practice, the h^{UV}_γ and c^{UV}_γ are iteratively solved on a separate 3D grid for each solvent atom type with $h^{VV}_{\alpha\gamma}$ being calculated in advance using Equation (9.26). Equation (9.34) is computed in reciprocal space while the closure relation is calculated in real space. This procedure is facilitated by the use of fast Fourier transforms (FFTs) to switch h^{UV} and c^{UV} between real and reciprocal space. At the end of each iteration the modified direct inversion in the iterative subspace (MDIIS) method [250] is applied to accelerate convergence and the residual tested against the predetermined tolerance. Upon convergence, thermodynamic properties of the solvation can be calculated via g^{UV} and the expression for the solvation free energy.

9.1.3.1 Solvation Free Energy

As with the 1D case, there exist analytical forms for the solvation free energy for both the HNC and KH closures [4, 250]. For HNC we have

$$\Delta\mu = \rho^V k_B T \sum_i \sum_\alpha \int \left[\frac{1}{2}\left(h^{UV}_{i\alpha}(\mathbf{r})\right)^2 - c^{UV}_{i\alpha}(\mathbf{r}) - \frac{1}{2}h^{UV}_{i\alpha}(\mathbf{r})c^{UV}_{i\alpha}(\mathbf{r})\right] d\mathbf{r} \tag{9.35}$$

and for KH we have

$$\Delta\mu = \rho^V k_B T \sum_i \sum_\alpha \int \left[\frac{1}{2}\left(h^{UV}_{i\alpha}(\mathbf{r})\right)^2 \Theta\left(-h^{UV}_{i\alpha}(\mathbf{r})\right) - c^{UV}_{i\alpha}(\mathbf{r}) - \frac{1}{2}h^{UV}_{i\alpha}(\mathbf{r})c^{UV}_{i\alpha}(\mathbf{r})\right] d\mathbf{r}. \tag{9.36}$$

9.1.3.2 Analytical Derivatives

Once an analytic expression for the solvation free energy has been derived, it is possible to calculate the derivate of this function [248]. Analytic derivatives allow 3D-RISM to be used in energy minimization routines [248] or in MD (see [244] and Chapter 10).

For the KH closure, this is done through differentiation of Equation (10.4) with respect to the positions of the solute atom, r_i. This gives the expression for the force on each atom:

$$\mathbf{f}^{UV}(\mathbf{r}_i) = -\frac{\partial\Delta\mu}{\partial\mathbf{r}_i} = -\sum_\alpha \rho_\alpha \int d\mathbf{R} g^{UV}_\alpha(\mathbf{R}) \frac{\partial u^{UV}_\alpha(\mathbf{R}-\mathbf{r}_i)}{\partial\mathbf{r}_i}. \tag{9.37}$$

9.2 Implementation

Little modification to either the standard Amber workflow (Fig. 9.1(a)) or the `SANDER` program (Fig. 9.1(b)) is required. In brief, to perform 3D-RISM-MD, the solvent direct and intramolecular pair correlation functions are required (see Equation (10.2)). These are generated by `RISM1D`, a stand alone implementation of 1D-RISM. The resulting solution for $\chi^{VV}_{\alpha\gamma}(r)$ is the only additional input file required for `SANDER`. The additional user input required by the 3D-RISM calculation is the grid dimensions and multiple time step (MTS) frequency.

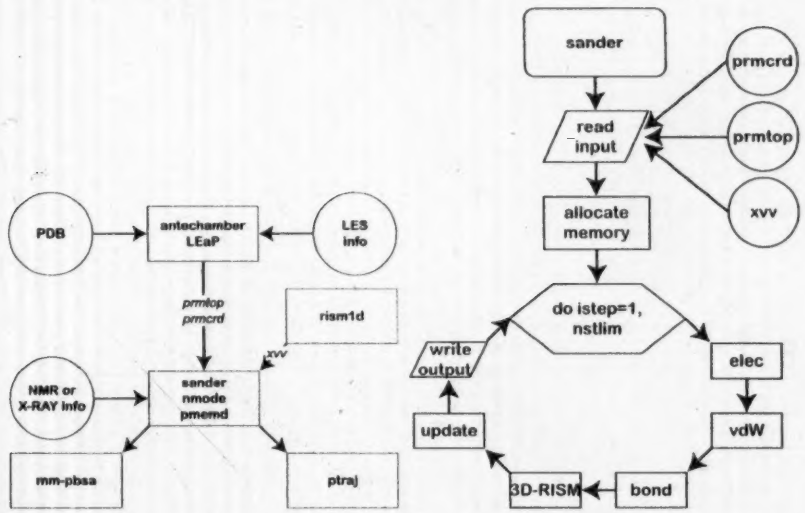


Figure 9.1: Flow charts for the (a) Amber package and (b)SANDER program. Necessary Modifications highlighted in red.

9.2.1 1D-RISM

As $\chi^{VV}_{\alpha\gamma}(r)$ need only be calculated once for a given solvent combination, density and temperature, 1D-RISM is implemented as a stand alone program, RISM1D. The only additional parameters that the user may wish to enter are the 1D grid size, temperature, number of molecular species in the mixture, and the dielectric constant of the mixture. For each species, the user must specify the number density and the input file containing the force field for the molecule. All standard solvents in Amber have been re-parameterized for 1D-RISM, including various water models, methanol, chloroform and various ions. That is, all hydrogens and dummy atoms have been given small Lennard-Jones radii to prevent unphysical overlap of charge sites and is detailed further in Section 9.3.1.

9.2.2 Implementation and Optimization of 3D-RISM

Modifications to the SANDER code base were minor. Other than calling the RISM3D subroutine, the only modifications were to add in calls for memory allocation and file input/output (Fig. 9.1).

Accurate solvation forces were produced with only modifications to adapt the 3D-RISM to the SANDER memory allocation model made. However, a single 3D-RISM calculation is roughly three orders of magnitude slower than a single time step for a system solvated with the same solvent model at the same volume and density. This is not unexpected as 3D-RISM performs a complete sampling of the solvent about the solute.

To obtain meaningful sampling of solute conformations it is necessary to optimize the 3D-RISM calculation. To achieve this goal three different optimization strategies were employed: 1) the pre- and post-processing of the solute-solvent potentials, long-range asymptotics and forces was accelerated using a cut-off scheme; 2) convergence of the 3D-RISM solution was

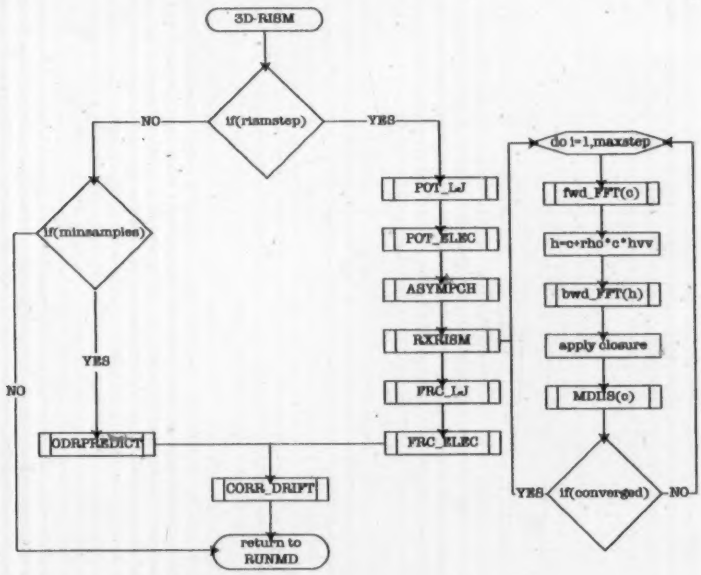


Figure 9.2: Flow chart for the 3D-RISM subroutine.

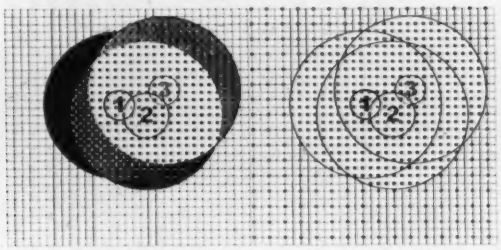
carried out in a minimal solvation box; 3) direct calculation of the 3D-RISM solvation forces was avoided altogether by interpolating current force based off of atom positions from previous time steps. The final structure of the `RISM3D` subroutine can be seen in Fig. 9.2.

The final structure of the `RISM3D` subroutine can be seen in Fig. 9.2.

9.2.2.1 Potential, Asymptotics and Force Calculations

Solute-solvent potential energy and force interactions and the asymptotic behavior of charges are computed on grids. Since the grid spacing is quite small (typically 0.25 Å) the number of interactions between the solute and the grid can significantly outnumber the interactions in an explicit solvent simulation. For example, alanine-dipeptide in a 32 Å × 32 Å × 32 Å box in pure SPC/E solvent requires $128^3 \times 22 \times 2 \times 2 \approx 1.8 \times 10^8$ interactions to calculate the potentials and forces for both solvent atom types, not including asymptotics (approximately $128^3 \times 22 \times 1.5 \approx 6.9 \times 10^7$ interactions). The same system with explicit solvent contains, in our example, 2416 atoms. Without cutoffs, this is $2416^2 \times 2 \approx 1.1 \times 10^7$ interactions for LJ and coulomb forces.

To reduce the total computational load of these calculations we employ a cutoff scheme adapted to the long-range properties of the LJ, coulomb and asymptotics functions. Even with these cutoffs in place it is necessary to assign a value to every grid point for the potential and asymptotics grids, though we can exclude some points when calculating the forces. The cutoff schemes described below allow us to assign constant values to the majority of grid points rather than calculating expensive functions for each one. The calculations are still $\mathcal{O}(M_x M_y M_z N^U)$ where M is the number of grid points per dimension and N^U is the number of solute atoms. However, because calculating the full interaction is an expensive sum, we are able to significantly reduce the computational cost (i.e. the coefficient of scaling) for the vast majority of the grid points.



(a) (b)

Figure 9.3: Cutoff schemes for solute-solvent (a) Lennard-Jones and (b) coulomb potentials and forces. There are three atoms in the system labeled '1' (blue), '2' (red) and '3' (green). Points that are directly used in the calculation are highlighted with large black dots. (a) For the Lennard-Jones interactions only points within each cutoff are used for each solute atom. Areas that include contributions from more than one atom are represented by mixed colours. (b) Coulomb interactions include contributions from all solute atoms. Inside the cutoff, all grid points are directly used. Outside the cutoff, a sparse grid is used directly (integration of forces) and is used to interpolate points on the fine grid (calculation of potential).

Potential Calculation The potential for the solute-solvent interactions is calculated as

$$u_{\alpha}^{UV}\left(\mathbf{R}_{abc}\right) = \sum_{i}^{N^{U}} \epsilon_{i\alpha}\left(\left(\frac{\sigma_{i\alpha}}{\mathcal{R}_{iabc}}\right)^{12} - 2\left(\frac{\sigma_{i\alpha}}{\mathcal{R}_{iabc}}\right)^{6}\right) + 332\frac{q_i q_\alpha}{\mathcal{R}_{iabc}} \tag{9.38}$$

where α is the solvent atom type, $\mathcal{R}_{iabc} = |\mathbf{R}_{abc} - \mathbf{r}_i|$, $\mathbf{R}_{abc}$ is the grid point at position (a, b, c) and $\mathbf{r}_i$ is the solute atom position. To reduce the computational load of this sum, we begin by identifying that the LJ and coulomb functions have different long-range behaviour and do the sum for each separately.

A simple truncation is the most straight forward approach to reducing the computational load of the LJ interaction. Thus, we adopt the same cutoff, r_{cut}, used for the LJ forces in the non-RISM parts of the calculation. This gives us

$$u_{\text{LJ},\alpha}^{UV}\left(\mathbf{R}_{abc}\right) = \sum_{i}^{N^{U}} \begin{cases} \epsilon_{i\alpha}\left(\left(\frac{\sigma_{i\alpha}}{\mathcal{R}_{iabc}}\right)^{12} - 2\left(\frac{\sigma_{i\alpha}}{\mathcal{R}_{iabc}}\right)^{6}\right) & \text{if } \mathcal{R}_{iabc} \leq r_{\text{cut}} \\ 0 & \text{if } \mathcal{R}_{iabc} > r_{\text{cut}} \end{cases} \tag{9.39}$$

As this is performed on a fixed grid we do not need to worry about cutoff lists and the distance only needs to be calculated for points that fall within the minimal bounding cube for the cutoff sphere. However, all grid points must still be initialized to zero so the scaling still depends on box size though to a much lesser extent. The schematic representation for this in 2D can be found in Fig. 9.3(a).

Coulomb interactions have a much longer tail than do LJ interactions, so a simple cutoff does not work in this case. We would like to have the contribution from every solute atom at every grid point. A reduction in the computational load can be achieved by recognizing that outside the collective cutoff volume of the solute, the electrostatic force decays in a smooth, predictable manner. Thus, we can explicitly calculate the contribution from all the solute atoms at a subset of the grid points and then use a fast interpolation method to calculate the contribution at the other grid sites. As the number of solute atoms increases this method

becomes more efficient. The adapted potential energy function looks like

$$
\begin{aligned}
u^{UV}_{\text{elec},\alpha}(\mathbf{R}_{abc}) = &\begin{cases} \sum_i^{N^U} 332\frac{q_i q_\alpha}{\mathcal{R}_{iabc}} & \text{if any } \mathcal{R}_{iabc} \le r_{\text{cut}} \\ 0 & \text{if every } \mathcal{R}_{iabc} > r_{\text{cut}} \end{cases} \\
&+ \begin{cases} \sum_i^{N^U} 332\frac{q_i q_\alpha}{\mathcal{R}_{iabc}} & \text{if every } \mathcal{R}_{iabc} > r_{\text{cut}} \text{ and } a, b \text{ and } c \text{ are even} \\ 0 & \text{if any } \mathcal{R}_{iabc} \le r_{\text{cut}} \end{cases} \\
&+ \begin{cases} I(u^{UV}_\alpha(\mathcal{R}_{i,a-1,b-1,c-1}), \\ \quad u^{UV}_\alpha(\mathcal{R}_{i,a+1,b-1,c-1}), \\ \quad u^{UV}_\alpha(\mathcal{R}_{i,a-1,b+1,c-1}), \\ \quad u^{UV}_\alpha(\mathcal{R}_{i,a+1,b+1,c-1}), \\ \quad u^{UV}_\alpha(\mathcal{R}_{i,a-1,b-1,c+1}), \\ \quad u^{UV}_\alpha(\mathcal{R}_{i,a+1,b-1,c+1}), \\ \quad u^{UV}_\alpha(\mathcal{R}_{i,a-1,b+1,c+1}), \\ \quad u^{UV}_\alpha(\mathcal{R}_{i,a+1,b+1,c+1})) & \text{if } u^{UV}_\alpha(\mathcal{R}_{i,a-1,b-1,c-1}) = 0 \\ 0 & \text{if } u^{UV}_\alpha(\mathcal{R}_{i,a-1,b-1,c-1}) \neq 0 \end{cases}
\end{aligned}
\tag{9.40}
$$

whre I is a fast polynomial interpolation subroutine [253] which uses values calculated in the first two terms. Note that the cutoff is applied only if grid point falls out of cutoff of *every* atom. In this way, the electrostatic contribution of every atom (or an approximation thereof) is applied to every grid point. This is illustrated in Fig. 9.3(b).

Force Calculation The solvation forces for each atom are integrated off the grids for the pair-distribution function, g^{UV}_α, of each solvent type

$$
\mathbf{f}^{UV}(\mathbf{r}_i) = \frac{\partial \Delta\mu}{\partial \mathbf{r}_i} = \sum_\alpha \rho_\alpha \int d\mathbf{r} g^{UV}_\alpha(\mathbf{R}_{abc}) \frac{\partial u^{UV}_\alpha(\mathcal{R}_{iabc})}{\partial \mathbf{r}_i}. \tag{9.41}
$$

where u_α is the position dependent potential energy of the solute and solvent atoms i and α. Expressed as a numerical sum, this becomes

$$
\begin{aligned}
\mathbf{f}^{UV}(\mathbf{r}_i) =& \frac{V}{M_x M_y M_z} \sum_\alpha \rho_\alpha \sum_a^{M_x} \sum_b^{M_y} \sum_c^{M_z} \frac{\partial u^{UV}_\alpha(\mathcal{R}_{iabc})}{\partial \mathbf{r}_i} g^{UV}_\alpha(\mathbf{R}_{abc}) \\
=& \frac{V}{M_x M_y M_z} \sum_\alpha \rho_\alpha \sum_a^{M_x} \sum_b^{M_y} \sum_c^{M_z} \\
& \left[12\epsilon_{i\alpha} \left(-\frac{\sigma^{12}_{i\alpha}}{\mathcal{R}^{14}_{iabc}} + \frac{\sigma^{6}_{i\alpha}}{\mathcal{R}^{8}_{iabc}} \right) - 332\frac{q_i q_\alpha}{\mathcal{R}^3_{iabc}} \right] \cdot \\
& \cdot (\mathcal{R}_{iabc,x}\hat{\imath} + \mathcal{R}_{iabc,y}\hat{\jmath} + \mathcal{R}_{iabc,z}\hat{k}) g^{UV}_\alpha(\mathbf{R}_{abc})
\end{aligned}
\tag{9.42}
$$

where V is the volume of the grid we are integrating. Though the expression for the forces does differ, how the contributions of the different force types decays with distance is still qualitatively the same. Thus, the same basic strategies for applying the cutoffs in used.

The expression for the LJ forces is then

$$\mathbf{f}_{\text{LJ}}^{UV}(\mathbf{r}_i) = \frac{V}{M_x M_y M_z} \sum_\alpha \rho_\alpha \sum_a^{M_x} \sum_b^{M_y} \sum_c^{M_z} \begin{cases} g_\alpha^{UV}(\mathbf{R}_{abc}) 12\epsilon_{i\alpha} \left(-\frac{\sigma_{i\alpha}^{12}}{\mathcal{R}_{iabc}^{14}} + \frac{\sigma_{i\alpha}^{6}}{\mathcal{R}_{iabc}^{8}} \right) & \text{if } \mathcal{R}_{iabc} \le r_{\text{cut}} \\ \cdot (\mathcal{R}_{iabc,x}\hat{\imath} + \mathcal{R}_{iabc,y}\hat{\jmath} + \mathcal{R}_{iabc,z}\hat{k}) & \\ 0 & \text{if } \mathcal{R}_{iabc}\ r_{\text{cut}} \end{cases} \tag{9.43}$$

analogous to Equation (9.39).

The procedure is slightly modified for the coulomb forces from Equation (9.40). Since we are integrating over the volume of the grid, we can avoid explicitly interpolating the sparse region beyond the cutoff by simply using larger integration step sizes. Also, for the sake of simplicity, rather than using a spherical cutoff we use the minimal cube that encloses the sphere. This gives the equation

$$\mathbf{f}_{\text{coulomb}}^{UV}(\mathbf{r}_i) = \frac{V}{M_x M_y M_z} \sum_\alpha \rho_\alpha \sum_a^{M_x} \sum_b^{M_y} \sum_c^{M_z} \begin{cases} -g_\alpha^{UV}(\mathbf{R}_{abc}) 332 \frac{q_i q_\alpha}{\mathcal{R}_{iabc}^3} & \text{if any } \mathcal{R}_{iabc} \le r_{\text{cut}} \\ \cdot (\mathcal{R}_{iabc,x}\hat{\imath} + \mathcal{R}_{iabc,y}\hat{\jmath} + \mathcal{R}_{iabc,z}\hat{k}) & \\ 0 & \text{if every } \mathcal{R}_{iabc}\ r_{\text{cut}} \end{cases}$$

$$+ \frac{8V}{M_x M_y M_z} \sum_\alpha \rho_\alpha \sum_{a=1,3,\ldots}^{M_x} \sum_{b=1,3,\ldots}^{M_y} \sum_{c=1,3,\ldots}^{M_z} \begin{cases} -g_\alpha^{UV}(\mathbf{R}_{a,b,c}) 332 \frac{q_i q_\alpha}{\mathcal{R}_{iabc}^3} & \text{if every } \mathcal{R}_{iabc} > r_{\text{cut}} \\ \cdot (\mathcal{R}_{iabc,x}\hat{\imath} + \mathcal{R}_{iabc,y}\hat{\jmath} + \mathcal{R}_{iabc,z}\hat{k}) & \\ 0 & \text{if any } \mathcal{R}_{iabc} \le r_{\text{cut}} \end{cases}$$

Asymptotics Correction of the long-range asymptotics is also calculated on a grid and is quite time consuming. At least the real space part of this correction can be optimized using a cutoff approach. The asymptotics for c^{UV} and h^{UV} in real space are give by

$$c_{\text{asm}}^{UV}(\mathbf{R}_{abc}) = \sum_i^{N^U} \begin{cases} \frac{q_i^U s}{2\sqrt{\pi}} & \text{if } \mathcal{R}_{iabc} = 0 \\ \frac{q_i^U}{\mathcal{R}_{iabc}} (1 - \text{erfc}(\mathcal{R}_{iabc}/s)) & \text{if } \mathcal{R}_{iabc} \neq 0 \end{cases} \tag{9.44}$$

$$h_{\text{asm}}^{UV}(\mathbf{R}_{abc}) = \sum_i^{N^U} \begin{cases} q_i^U h_{\text{coef}}(\mathcal{R}_{iabc}, \kappa, s) & \text{if } \mathcal{R}_{iabc} = 0 \\ \frac{q_i}{\mathcal{R}_{iabc}} h_{\text{coef}}(\mathcal{R}_{iabc}, \kappa, s) & \text{if } \mathcal{R}_{iabc} \neq 0 \end{cases} \tag{9.45}$$

where $\text{erfc}(x)$ is the complementary error function, κ is the inverse Debye length, s is the smear applied to the coulomb potential and

$$h_{\text{coef}}(\mathcal{R}_{iabc}, \kappa, s) = e^{-(\kappa s/2)^2} \left[e^{-\kappa \mathcal{R}_{iabc}} \text{erfc}\left(\frac{\kappa s}{2} - \frac{\kappa}{s} \right) - e^{\kappa \mathcal{R}_{iabc}} \text{erfc}\left(\frac{\kappa s}{2} + \frac{\kappa}{s} \right) \right] / 2 \tag{9.46}$$

As $\mathcal{R}_{iabc} \to \infty$ we quickly approach the limits

$$\lim_{\mathcal{R}_{iabc} \to \infty} (1 - \text{erfc}(\mathcal{R}_{iabc}/s) = 1 \tag{9.47}$$

$$\lim_{\mathcal{R}_{iabc} \to \infty} h_{\text{coef}}(\mathcal{R}_{iabc}, \kappa, s) = e^{-(\kappa s/2)^2} \tag{9.48}$$

The cutoff can then be calculated at run time to produce arbitrarily small errors. For typical values of $s = 1$ and $\kappa = 0$ we can achieve an error threshold of $< 10^{-7}$ with a cutoff of

$r_{\text{cut}} = 3.9$ Å. This gives the following formulas

$$c^{UV}_{\text{asm}}(\mathbf{R}_{abc}) = \sum_i^{N^U} \begin{cases} \frac{q_i^U s}{2\sqrt{\pi}} & \text{if } \mathcal{R}_{iabc} = 0 \\ \frac{q_i^U}{\mathcal{R}_{iabc}}(1 - \text{erfc}(\mathcal{R}_{iabc}/s)) & \text{if } \mathcal{R}_{iabc} \neq 0 \\ \frac{q_i^U}{\mathcal{R}_{iabc}}(1 - \text{erfc}(r_{\text{cut}}/s)) & \text{if } \mathcal{R}_{iabc} > r_{\text{cut}} \end{cases} \tag{9.49}$$

$$h^{UV}_{\text{asm}}(\mathbf{R}_{abc}) = \sum_i^{N^U} \begin{cases} q_i^U h_{\text{coef}}(\mathcal{R}_{iabc}, \kappa, s) & \text{if } \mathcal{R}_{iabc} = 0 \\ \frac{q_i^U}{\mathcal{R}_{iabc}} h_{\text{coef}}(\mathcal{R}_{iabc}, \kappa, s) & \text{if } \mathcal{R}_{iabc} \neq 0 \\ \frac{q_i^U}{\mathcal{R}_{iabc}} h_{\text{coef}}(r_{\text{cut}}, \kappa, s) & \text{if } \mathcal{R}_{iabc} > r_{\text{cut}} \end{cases} \tag{9.50}$$

Since $(1 - \text{erfc}(r_{\text{cut}}/s))$ and $h_{\text{coef}}(r_{\text{cut}}, \kappa, s)$ can be calculated in advance, these otherwise expensive calculations can be avoided.

9.2.3 3D-RISM Convergence

The bulk of calculation time is spent iteratively solving the 3D-RISM equations. Some straightforward methods can be used to speed up this calculation, such as modified direct inversion in the iterative subspace (MDIIS) [250, 254], using an efficient fast Fourier transform, like FFTW [255], and using a previous solution as an initial guess [244]. While we employ all of these methods we also make further optimizations by using a solution propagator and solving the 3D-RISM equation for a minimal solvation box rather than the complete super-cell.

9.2.3.1 Solution Propagation

Using the solution from the previous time step for the initial guess for the current time step greatly reduces the number of iterations required to converge upon a solution. However, even very small changes in the positions of atoms can require several iterations to converge to a new solution. The number of iterations required can be further reduced by propagating several past solutions to form an initial guess. Both linear,

$$c_i = 2c_{i-1} - c_{i-2}, \tag{9.51}$$

and quadratic,

$$c_i = 3c_{i-1} - 3c_{i-2} + c_{i-3}, \tag{9.52}$$

forms improve the quality of the initial guess but higher orders do not.

9.2.4 Multiple-Time Step Algorithms

With a target linear grid spacing of 0.25 Å for 3D-RISM, compared to the target of 1 Å for Ewald summation methods, it is clear that even if our solution converges in one iteration it will be considerably slower than using explicit solvent. Clearly, for 3D-RISM to be an effective implicit solvent for MD a multiple time step (MTS) algorithm must be used to avoid solving 3D-RISM on each time step.

Typically, MTS algorithms are used for expensive long-range electrostatics calculations that vary slowly with time. Two common methods that have been used for this are impulse [119] and extrapolative MTS [96]. Both have been implemented for 3D-RISM-MD and are reviewed below. We also suggest a third method, interpolative MTS.

9.2.4.1 Impulse Based MTS

The most common MTS algorithms currently in use are impulse based MTS schemes derived from using the Trotter factorization of the Liouville operator. For a derivation of three-class and two-class algorithms see Schlick [96] and Frenkel and Smit [119] respectively.

The basic premise of this method is splitting the force into different terms based on their characteristic frequencies. In a three-class method these would be short (e.g. bonded interactions), medium (e.g. short-range non-bonded) and long (e.g. long-range electrostatics) interactions. Time steps Δt_1, Δt_2 and Δt_3 are defined for each of the interactions with

$$\Delta t_3 = k_2 \Delta t_2 = k_2 k_1 \Delta t_1 \tag{9.53}$$

where k_1 and k_2 are integers. At k_1 and $k_1 k_2$ time steps the medium and long forces are calculated and the appropriate impulse is applied.

Algorithm 1: Three-class impulse MTS based on velocity Verlet

for istep = 1 to NSTEP
  $F_{\text{long}} = -\nabla E_{\text{long}}(X)/M$
  $V \leftarrow V + \frac{\Delta t_3}{2} F_{\text{long}}$
  for $j = 1$ to k_2
    $F_{\text{med}} = -\nabla E_{\text{med}}(X)/M$
    $V \leftarrow V + \frac{\Delta t_2}{2} F_{\text{med}}$
    for $i = 1$ to k_1
      $F_{\text{short}} = -\nabla E_{\text{short}}(X)/M$
      $V \leftarrow V + \frac{\Delta t_1}{2} F_{\text{short}}$
      $X \leftarrow X + \Delta t_1 V$
      $V \leftarrow V + \frac{\Delta t_1}{2} F_{\text{short}}$
    endfor
    $V \leftarrow V + \frac{\Delta t_2}{2} F_{\text{med}}$
  endfor
  $V \leftarrow V + \frac{\Delta t_3}{2} F_{\text{long}}$
endfor

These algorithms have the same attractive qualities as the Verlet integrator from which they originate. That is, they are symplectic and time-reversible. However, they are susceptible to resonance artifacts. As the largest time step approaches the shortest period, or half-period, of the system, resonance occurs and the simulation becomes unstable. For simple systems, for example the 1D harmonic oscillator, stability can be recovered with a larger outer time step that is not an integer multiple of the resonance frequency. In practice, this does not work for complex biological molecules, likely due to the wide range of frequencies present in the system. For this reason, impulse MTS algorithms are practically limited to a 4-5 fs large time step.

9.2.4.2 Extrapolative MTS

As in impulse MTS, slow and medium forces are calculated every $k_2 k_1$ and k_1 time steps. Rather than applying these forces in single large impulses, the force is applied to every inner time step.

Algorithm 2: Extrapolative MTS based on position Verlet

for istep = 1 to NSTEP
  $X^* \leftarrow X + \frac{\Delta t_2}{2} V$
  $F_{\text{long}} = -\nabla E_{\text{long}}(X^*)/M$

for $j = 1$ to k_2
  $X^* \leftarrow X + \frac{\Delta t_2}{2} V$
  $F_{\text{med}} = -\nabla E_{\text{med}}(X^*)/M$
  for $i = 1$ to k_1
    $X \leftarrow X + \frac{\Delta t_1}{2} V$
    $F_{\text{short}} = -\nabla E_{\text{short}}(X)/M$
    $V \leftarrow V + \frac{\Delta t_1}{2}(F_{\text{short}} + F_{\text{med}} + F_{\text{long}})$
    $X \leftarrow X + \frac{\Delta t_1}{2} V$
  endfor
endfor
endfor

Note that midpoint extrapolation is used for both the long and midrange forces. That is, F_{long} and F_{med} are evaluated at X^* and not X. Using $X^* \leftarrow X$ gives constant extrapolation.

Unfortunately, such method are non-symplectic and result in a net energy drift.

Langevin Dynamics and LN Problems with energy drift and resonance artifacts for extrapolative and impulse MTS methods can be, to some degree, overcome by introducing a stochastic term in the form of Langevin dynamics. For impulse methods the improvement is not profound. Resonance artifacts still occur though their magnitude is reduced.

The LN method has been developed for extrapolative MTS with much improved results. With PME a stable outer time step of 12-16 fs has been achieved [96]. Direct sum electrostatics allow time steps of 48 fs for long-range contributions (> 6.5Å) [96]. While several variants have been developed, the direct force version is the most practical and is quite similar to Algorithm 2.

Algorithm 3: LN MTS

for istep = 1 to NSTEP
  $F_{\text{long}} = -\nabla E_{\text{long}}(X)$
  for $j = 1$ to k_2
    $X^* \leftarrow X + \frac{\Delta t_2}{2} V$
    $F_{\text{med}} = -\nabla E_{\text{med}}(X^*)$
    for $i = 1$ to k_1
      $X \leftarrow X + \frac{\Delta t_1}{2} V$
      $F_{\text{short}} = -\nabla E_{\text{short}}(X)$
      $V \leftarrow V + \frac{\Delta t_1}{M}(F_{\text{short}} + F_{\text{med}} + F_{\text{long}} + R)/(1 + \gamma \Delta t_1)$
      $X \leftarrow X + \frac{\Delta t_1}{2} V$
    endfor
  endfor
endfor

R and γ are due to use the Langevin equation are discussed in Section 9.2.4.3. Here midpoint extrapolation is used only for the midrange forces. That is, F_{med} are evaluated at X^* and not X.

9.2.4.3 MTS in Amber

When using Langevin dynamics in Amber the LBP integrator is used[256].

Algorithm 4: Langevin dynamics in Amber

for istep = 1 to NSTEP
  $F = -\nabla E(X)$

$V \leftarrow \left(V(1-\gamma\Delta t/2) + \frac{\Delta t}{M}(F+R)\right)/(1-\gamma\Delta t/2)$
$X =\leftarrow X + V\Delta t$
endfor

where V and X are the velocity and position of the particle, Δt is the size of the time step, $\gamma = \eta/m$ where η is the friction constant. R is generated by a pseudo random number generator such that

$$< R_n^2 >= 2M\gamma kT/\Delta t \tag{9.54}$$

Combining the LBP integrator with force-splitting, impulse MTS is straightforward. Algorithm 4 becomes

Algorithm 5: Impulse MTS with Langevin dynamics in Amber

for istep = 1 to NSTEP
$F_{\text{long}} = -\nabla E_{\text{long}}(X)$
for $i = 1$ to k_1
$F_{\text{short}} = -\nabla E_{\text{short}}(X)$
if $i == 1$ **then**
$V \leftarrow \left(V(1-\gamma\Delta t_1/2) + \frac{\Delta t_1}{M}(F_{\text{short}} + F_{\text{long}}k_1 + R)\right)/(1-\gamma\Delta t_1/2)$
else
$V \leftarrow \left(V(1-\gamma\Delta t_1/2) + \frac{\Delta t_1}{M}(F_{\text{short}} + R)\right)/(1-\gamma\Delta t_1/2)$
endif
$X \leftarrow X + V\Delta t_1$
endfor
endfor

9.2.4.4 Modified MTS for 3D-RISM in Amber

While the solvation forces that 3D-RISM calculates are both short and long range, the forces vary slowly in time due to their mean-field nature. It should be then possible to use time steps for 3D-RISM that are considerably longer than the 1-2 fs commonly used for constrained MD. However, 3D-RISM is still subject to the same resonance artifacts that other long range forces are when using MTS methods.

As 3D-RISM is an implicit solvation method that requires very long time steps for practical use in MD, LN becomes an attractive choice. This combines the desirable properties of Langevin dynamics for constant temperature simulations with implicit solvents with the long time step capabilities of LN.

LN was originally developed as a three-class force-splitting method. While this can be used for 3D-RISM/MD, splitting fast and intermediate forces offers little speedup compared to the cost of 3D-RISM. Therefore, we use a two-class method with the same extrapolative MTS as used in the outer loop of LN with the current Langevin integrator found in Amber.

Algorithm 6: Two-class LN MTS in Amber

for istep = 1 to NSTEP
$F_{\text{3D-RISM}} = -\nabla E_{\text{3D-RISM}}(X)$
for $i = 1$ to k_1
$X \leftarrow X + \frac{\Delta t_1}{2}V$
$F_{\text{short}} = -\nabla E_{\text{short}}(X)$
$V \leftarrow V + \frac{\Delta t_1}{M}(F_{\text{short}} + F_{\text{3D-RISM}} + R)/(1+\gamma\Delta t_1)$
$X \leftarrow X + \frac{\Delta t_1}{2}V$
endfor
endfor

9.2.4.5 Coordinate Based Interpolative MTS

Solvation forces computed by 3D-RISM contain both fast and slow varying contributions that cannot be split. Even if LN-MTS can get us past the resonance limit of impulse MTS we are still limited by the period of the fastest contributions to the solvation forces.

Ideally, we would compute a full 3D-RISM solution once for several time steps and have smoothly varying forces applied in between that are reasonably close to the true 3D-RISM solvation forces. In principle, this can be done by creating a analytical model relating coordinate positions and the solvation forces experienced by the solute atoms. A series of 3D-RISM calculations can be done for a number of time steps to collect force-position data to be used to fit the model. While not giving the true forces, such a method can give approximate forces that are arbitrarily close to the true forces if a suitable model is provided with enough data for a proper fit.

Before the minimum number data points have been collected through explicit 3D-RISM force calculations, the calculation may proceed according to Algorithm 5 or 6. Once the sampling phase is complete, the following Algorithm is invoked

Algorithm 7: Interpolation MTS in Amber

for istep = 1 to NSTEP
  $F_{\text{3D-RISM}} = -\nabla E_{\text{3D-RISM}}(X)$
  `store_data`($F_{\text{3D-RISM,old}}$, X_{old}, $F_{\text{3D-RISM}}$, X)
  `fit_model`($F_{\text{3D-RISM,old}}$, X_{old})
  for $i = 1$ to k_1
    $X \leftarrow X + \frac{\Delta t_1}{2} V$
    $F_{\text{short}} = -\nabla E_{\text{short}}(X)$
    $F_{\text{3D-RISM}}$ = `predict_3D_RISM_forces`(X)
    $V \leftarrow V + \frac{\Delta t_1}{M}(F_{\text{short}} + F_{\text{3D-RISM}} + R)/(1 + \gamma \Delta t_1)$
    $X \leftarrow X + \frac{\Delta t_1}{2} V$
  endfor
endfor

Three subroutines are invoked to handle the modelling of the forces, `store_data`, `fit_model` and `predict_3D_RISM_forces`. Note that *only* predicted forces are used and not the direct 3D-RISM forces. This is to prevent possible discontinuity in the forces but means that the solvation energy does not directly correspond to the solvation forces. This is not a problem as the solvation energies are only used in the context of replica exchange MD (REMD). Even in this context, REMD simulations using explicit solvent for forces and GB for the solvation energy have been successful [147].

A particularly simple, but effective model, is

$$f_{\gamma,\Lambda} = \eta_0 + \eta_1(r_{1,x} - r_{\gamma,x}) + \eta_2(r_{1,y} - r_{\gamma,y}) + \eta_3(r_{1,z} - r_{\gamma,z}) + \cdots + \eta_{3n-2}(r_{n,x} - r_{\gamma,x}) + \eta_{3n-1}(r_{n,y} - r_{\gamma,y}) + \eta_{3n}(r_{n,z} - r_{\gamma,z}) \quad (9.55)$$

where λ is the x, y or z component and η_0 to η_{3n} are the parameters of the model to be fit. $n < N_{sol}$ and may be the number of atoms within a cutoff radius of atom γ, giving a total of $3n + 1$ parameters for each component of the force for solute atom. While better models can be devised, this is sufficient to show the potential of the method.

Of course, when designing a model certain considerations must be made. The number of free parameters should be limited if possible. The larger the number of parameters the larger the number of 3D-RISM solutions required to accurately fit it. Also, a model that is fast to compute accelerates both the fitting procedure and the calculation of forces.

Model	ϵ kcal/mol	σ Å
Minimal H	0.04600	0.22449
TIP3P Large H	0.01520	0.69388
SPC/E Large H	0.01553	0.65424

Table 9.1: Hydrogen radii for water and 3D-RISM. Minimal H may be used with either TIP3P or SPC/E. ϵ for TIP3P Large and SPC/E is chosen at 10% of the ϵ value for oxygen and the radii coincide with the oxygen radii.

9.3 Free Energy of Solvent Polarization

In MD simulations, the principal interest of using 3D-RISM is calculating the free energy of solvation and the solvation forces derived from it. Accurate solvation free energies are determined by the quality of the 3D-RISM calculation and the parameters of the solvent model being used. Ideally, we would like to reproduce experimentally determined solvation free energies. However, as we are using an explicit solvent model as input for 3D-RISM we should first reproduce the behaviour of the explicit model. If we are successful in this, improvements to the underlying model compared to experiment, in principle, will also lead to better agreement between 3D-RISM and experiment.

9.3.1 Parameters for Site-Site Water Models

A major issue for 1D-RISM and, to lesser extent, 3D-RISM, is solvent atoms buried within the LJ radius of other solvent atoms. The most relevant example is that of explicit water models. As discussed in Section 4.3.1 and illustrated in Figure 4.5, hydrogens are encompassed within the LJ radius of their bonded oxygen and have no LJ radius of their own. For MD studies, this is not a problem; the stiff LJ potential of the oxygen prevents catastrophic charge overlap between water hydrogens and any other atom [2].

Approximations made in the closure, notably neglecting the bridge function, allow for the catastrophic collapse of solvent or solute sites onto the water model's hydrogen sites. In explicit MD simulations, the hydrogen sites are buried within the radius of the oxygen of the water model and such site-site overlaps are not possible. To prevent this collapse and allow 1D-RISM to converge to a solution, the strategy of giving LJ parameters to the hydrogens has been adopted and several different parameter sets have been used in the past [257–260]. In this work we explore a minimal radius [261] versus a radius that coincides with the oxygen radius along the O-H bond vector as in (see Figure 9.5). Example parameters for Equation (5.44) are given in Table 9.1. Parameters for a minimal radius are known to preserve the second peak observed in the experimental measured $g_{OO}(r)$ as in Figure 9.4. The alternative, large radius approach is used in this thesis. These parameters are chosen to reproduce the effects of the oxygen radius around the hydrogens and provide a better fit with the solvation free energies of the models they are based on.

9.3.2 Free Energy of Solvent Polarization

The best way to quantify the accuracy of 3D-RISM is by comparing solvation free energies from 3D-RISM against MD calculations utilizing that model directly. Calculating such solva-

[2] It is worth noting that, starting with the CHARMM22 version, the CHARMM forcefield was specifically optimized for a modified TIP3P water model with small LJ radii on the hydrogens [120].

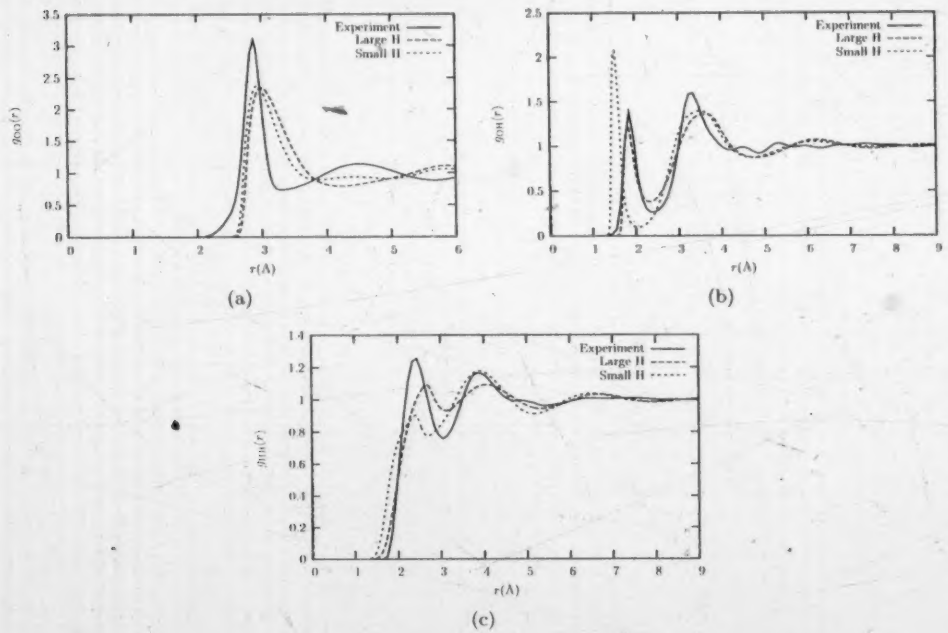


Figure 9.4: RDFs from experiment [71] and 1D-RISM with the SPC/E model.

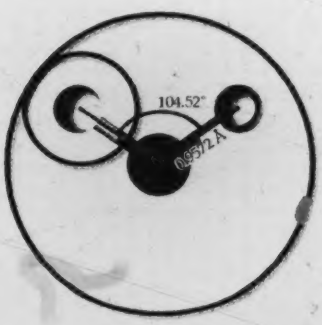


Figure 9.5: SPC/E with alternate hydrogen Lennard-Jones parameters. The oxygen LJ radius is the largest circle, the smallest is the hydrogen radius as in [261] while second largest represents the large hydrogen radius used in this work.

tion free energies with explicit solvent typically utilizes thermodynamics integration (TI), a computationally demanding procedure. To quantify the quality of Poisson equation solutions (PE) and various generalized Born (GB) models, Roe *et al.* [77] used TI on only the partial charges of ALA_{10}, thus calculating the free energy of solvent polarization. This provides the opportunity to directly compare 3D-RISM against TIP3P, PE and GB models.

9.3.2.1 Methods

The average free energy of solvent polarization (ΔG_{pol}) was calculated for four conformations of ALA_{10} (Ace-A_{10}-NH_2): α-helix (alpha), left-handed α-helix (left), polyproline II helix (PP2) and β-hairpin (hairpin). The generation of the structures used for each of these conformations is given in Roe *et al.* [77]. In brief, the four structures were created with the backbone ϕ/ψ angle values: alpha = -57.8°,-47.0°, left = 57.8°,47.0°, PP2 = -75.0°,145.0° and hairpin was taken from the backbone of Trpzip2 (PDB ID 1LE1) [262]. These structures were maintained during thermodynamic integration (TI) with harmonic position restraints while the atomic point charges where increased from 0 to their ff99SB values both in the presence and absence of TIP3P [69] water. Further details of this calculation can be found in [77].

3D-RISM solvation free energies (ΔG_{sol}) were calculated for each of the 1000 structures of each conformation from the TI calculation using both SPC/E and TIP3P water models with hydrogen LJ parameters given in Table 9.1. In total, six sets of 3D-RISM solvation free energies were calculated with different parameters and precisions. Calculations at grid resolution of 0.5 Å and residual tolerance of 10^{-3} were done for SPC/E and TIP3P with small and large hydrogen radii. Based off these results, a grid resolution of 0.25 Å and 10^{-5} was used for only SPC/E and TIP3P. As 3D-RISM calculates the full solvation free energy, ΔG_{pol} was calculated as the difference between ΔG_{sol} for ALA_{10} with and without point charges. To facilitate the computation, the principal axis for each structure was oriented along the z-axis, permitting the smallest possible solvation box. The same solvation 36 Å × 36 Å × 60 Å box was used for all calculations, giving a minimum distance of 14 Å from the solute to the edge of the box.

9.3.2.2 Results and Discussion

Table 9.2 gives the 3D-RISM polarization free energies of the various conformations of ALA_{10}. There is qualitative agreement between all of the figures and the ordering of the conformations (PP2 < hairpin < left < alpha). Differences in ΔG_{pol} ($\Delta\Delta G_{\text{pol}}$) also generally agree. However, there is a large discrepancy between values for large and small hydrogen radii for ΔG_{pol}.

Standing out from Table 9.2 is the qualitative agreement between the large hydrogen models for the two very different tolerances and resolutions used. All values agree within one standard deviation, agreeing within 1% or better. There is, however, roughly a 25 fold difference in the computation time between the two sets of calculations.

The solvation free energies are also in agreement with experiment (see Section 4.2.1.2) in that 3D-RISM gives a positive solvation free energy for all conformations (data not shown). As we have no other figures to compare against, this qualitative comparison will have to suffice.

The highest precision 3D-RISM figures, along with results for explicit TIP3P, PE and GB [77], are given in Table 9.3. We observe excellent quantitative agreement between both 3D-RISM and TIP3P TI. As expected, the TIP3P 3D-RISM calculations are in better agreement with TI than SPC/E due to the shared model parameters. In fact, the two agree within one standard deviation of TIP3P-3D-RISM results and 2-3 standard errors of the of the TI calculations.

	Minimal Radii		Coincident Radii			
	10^{-3}, 0.5 Å		10^{-3}, 0.5 Å		10^{-5}, 0.25 Å	
	SPC/E	TIP3P	SPC/E	TIP3P	SPC/E	TIP3P
			(A) ΔG_{pol}			
PP2	-94.61±1.07	-93.60±1.07	-79.19±0.98	-76.89±0.96	-79.12±1.28	-76.82±1.27
alpha	-56.36±0.94	-55.79±0.93	-46.23±0.84	-44.88±0.82	-46.25±1.37	-44.91±1.31
left	-62.35±1.04	-61.81±1.03	-52.81±0.93	-51.44±0.92	-52.92±0.95	-51.60±1.22
hairpin	-70.12±1.32	-69.36±1.31	-57.92±1.18	-56.08±1.17	-57.78±1.49	-56.00±1.17
			(B) $\Delta\Delta G_{pol}$			
PP2-alpha	-38.26	-37.81	-32.96	-32.01	-32.87	-31.91
PP2-left	-32.26	-31.79	-26.38	-25.45	-26.19	-25.22
PP2-hairpin	-24.49	-24.24	-21.27	-20.81	-21.33	-20.82
alpha-left	6.00	6.02	6.58	6.56	6.67	6.69
alpha-hairpin	13.76	13.57	11.69	11.21	11.53	11.09
left-hairpin	7.76	7.55	5.11	4.65	4.86	4.40

Table 9.2: 3D-RISM ΔG_{pol} of ALA_{10} using various parameters and resolutions. Errors are one standard deviation from the mean.

The main discrepancy between the 3D-RISM and the TI calculations is the hairpin conformation. This is also the calculation that has the largest error of all the TI calculations, indicating that this may be a particularly troublesome conformation. Both SPC/E and TIP3P calculations, however, still fair better overall for this conformation than the implicit models.

9.4 Conclusions

3D-RISM has been implemented in MD package Amber. This has specifically included optimizing both 3D-RISM and the coupling of 3D-RISM with MD (Section 9.2 and Chapter 10) and the characterization and validation of 3D-RISM output (Section 9.3 and Chapter 10).

Computational efficiencies introduced to 3D-RISM include streamlining the grid based calculations of solution potential, electrostatic asymptotics and solvation forces using a cutoff method. Care was taken in all cases to preserve the long range behaviour of the calculations. For electrostatics, in particular, the long range behaviour of the potential was accommodated by calculating the potential on a sparse grid, outside the cutoff radius, and interpolating the potential for grid points in between. For the electrostatic forces, the two grid approach was utilized again by integrating a small grid, based on a cutoff, at full resolution and a large grid, outside the cutoff, at half the resolution.

It should be noted that the scaling order of computations is not reduced, as all grid points must still be visited. Rather, the number of grid points requiring expensive calculations is substantially reduced. For Alanine-dipeptide with a 0.5 Å grid spacing, this gives up to a 3 fold speedup for these calculations. Smaller grid spacing and larger solutes should take better advantage of these optimizations.

Optimizations that take advantage of the highly correlated structures produced in MD are aimed at reducing the number of iterations required for 3D-RISM and for avoiding 3D-RISM calculations altogether.

Using the solution from the previous 3D-RISM calculation as an initial guess is a trivial method to reduce the number of iterations required to converge the new solution. By propagating the solutions from up to three previous solutions, an better initial guess is obtained and even fewer iterations are required

3D-RISM is still too expensive to execute every time step and it is best to avoid the calculation if possible. MTS methods are often used to accomplish such goals. However, the energy conserving impulse MTS scheme can not be extended beyond 4-5 fs time steps due to

	TIP3P	3D-RISM		PE	GBHCT	GBOBC	GBNeck
		TIP3P	SPC/E				
			(A) ΔG_{pol}				
alpha	-44.08±0.04	-44.91±1.27	-46.25±1.37	-47.97±0.77	-51.69±1.21	-49.38±1.21	-43.26±0.90
PP2	-76.39±0.15	-76.82±1.31	-79.12±1.28	-78.05±0.91	-77.35±1.05	-78.07±1.09	-77.59±1.02
left	-51.30±0.12	-51.60±1.22	-52.92±0.95	-54.85±0.90	-55.05±1.08	-52.67±1.10	-48.19±0.91
hairpin	-54.16±0.25	-56.00±1.17	-57.78±1.49	-57.28±1.13	-57.48±1.45	-56.03±1.47	-52.85±1.29
			(B) $\Delta\Delta G_{\text{pol}}$				
PP2-alpha	-32.31	-31.91	-32.87	-30.07	-25.67	-28.69	-34.33
PP2-left	-25.09	-25.22	-26.19	-23.19	-22.31	-25.40	-29.40
PP2-hairpin	-22.23	-20.82	-21.33	-20.77	-19.87	-22.03	-24.73
alpha-left	7.22	6.69	6.67	6.88	3.36	3.29	4.93
alpha-hairpin	10.08	11.09	11.53	9.31	5.80	6.66	9.60
left-hairpin	2.86	4.40	4.86	2.43	2.43	3.37	4.67
			(C) $\Delta\Delta G_{\text{pol}}$ Root-Mean-Square Deviations				
overall		0.99	1.21	1.39	3.89	2.60	2.51
PP2		0.85	0.88	1.89	4.37	2.10	3.11
non-PP2		1.11	1.46	0.55	3.34	3.02	1.71
hairpin		1.34	1.52	1.53	2.83	2.00	1.80
non-hairpin		0.39	0.78	1.58	4.72	3.09	3.05

Table 9.3: TI, 3D-RISM, PE and GB ΔG_{pol} of ALA_{10}. The two highest precision 3D-RISM calculations using the large hydrogen LJ parameters are given in this table. Errors are given as standard deviations except for the TIP3P TI calculations. TIP3P TI, PE and GB data are from Roe *et al.* [77].

resonance artifacts. LN extrapolative MTS has been used for other slow time scale interactions, such as long range electrostatics, to move beyond this resonance limit. Unfortunately, fast and slow time scale solvation forces cannot be decoupled in 3D-RISM so LN MTS offers little advantage. By using a simple model to predict the forces based on the conformation of the solute, the size of the 3D-RISM time step can be extended to 100s of femtoseconds. This comes at the cost of heating the system, so a relatively large friction coefficient must be used with Langevin dynamics. True dynamics are already lost by the use of a mean-field solvation model so the cost of losing accurate dynamics is small compared to the gains in sampling.

Finally, we have characterized the performance of 3D-RISM using metrics important for MD. Namely, we have shown that 3D-RISM is able to accurately reproduce the free energy of solvent polarization associated with the underlying explicit model. This is of critical importance for differentiating, and correctly sampling, biomolecules conformations. We have also shown that for 3D-RISM coupled with MD the total energy drift and net force can be made arbitrarily small. This does come at a computational cost however.

This work is simply the first step in coupling 3D-RISM and MD. In the immediate future, more work needs to be done on validating and accelerating the combination.

Validation will come in form of extensive MD runs on larger biomolecules and by further demonstrating the accuracy of full solvation free energies.

Further speed enhancements will come in three areas. Parallelization will be necessary for the method to be practical but will not enhance the method relative to current explicit or implicit models. An improved interpolative model for MTS will improve predictions, lower the magnitude of the friction coefficient necessary and, possibly, allow larger time steps to be used. Finally, we can take advantage of free boundary conditions by aligning the principal axis of the solute with the coordinate axes and setting the minimal solvent box to accommodate it. The computational gains for globular proteins may be modest but elongated proteins could gain substantially.

Chapter 10

Three-dimensional molecular theory of solvation coupled with molecular dynamics in Amber[1]

10.1 Introduction

Molecular dynamics (MD) simulation with explicit solvent, in particular, available in the Amber molecular dynamics package [5], yields accurate and detailed modeling of biomolecules (e.g. proteins and DNA) in solution, provided the processes to be described are within accessible time scales, typically up to tens of nanoseconds. A major computational burden comes from the treatment of solvent molecules (usually water, sometimes cosolvent, and counterions/buffer or salt for electrolyte solutions) which typically constitute a large part of the system. Moreover, solvent enters pockets and inner cavities of the proteins through their conformational changes, which is a very slow process and nearly as difficult to model as protein folding. Of no surprise, then, is the considerable interest in MD simulation with solvent degrees of freedom contracted by using implicit solvation approaches. In particular, the generalized Born (GB) model [97], in which the solvent polarization effects are represented by a cavity in dielectric continuum (optionally, with Debye screening by the charge distribution of structureless ions in the form of the Yukawa screened potential), whereas the non-electrostatic contributions are phenomenologically parameterized against the solvent accessible area and excluded volume of the biomolecule. The cavity shape is formed by rolling a spherical probe, of a size to be parameterized for each solvent, over the surface of the biomolecule. The polarization energy follows from the solution to the Poisson equation, which is computationally expensive, and is approximated in the GB model for fast calculation by algebraic expressions interpolating between the simple cases of two point charges in a spherical cavity. Conceptually transparent and computationally simple, the GB model has long been popular, including its implementations in the Amber molecular dynamics package [5]. However, it bears the fundamental drawbacks of implicit solvation methods: the energy contribution from solvation shell features such as hydrogen bonding can be parameterized but not represented in a transferable manner; the three-dimensional variations of the solvation structure, in particular, the second solvation shell are lost; the volumetric

[1] A version of this chapter has been submitted for publication.
T. Luchko, S. Gusarov, D. A. Case, J. Tuszynski and A. Kovalenko. Three-dimensional molecular theory of solvation coupled with molecular dynamics in Amber. Phys. Chem. Chem. Phys. Submitted 2008/04/01.

properties of the solute are not well defined; the non-electrostatic solvation energy terms are empirically parameterized, and therefore, effective interactions like hydrophobic interaction and hydrophobic attraction are not described from first principles and thus are not transferable to new systems with complex compositions (e.g. with cosolvent and/or different buffer ions); the entropic term is absent in continuum solvation, thus excluding from consideration all associated effects, for example the energy-entropy balance for the temperature control over supramolecular self-assembly in solution. To this end, the notion of a surface accessible surface, defined as that delineated by the center of the probe "rolled" over the surface becomes meaningless for inner cavities of biomolecules hosting just a few solvent molecules.

An attractive alternative to continuum solvation is the three-dimensional molecular theory of solvation, also known as the 3D reference interaction site model (3D-RISM) [250]. Starting from the force field, it operates with solvent distributions rather than individual molecules, but yields the solvation structure and thermodynamics from first principles of statistical mechanics. It properly accounts for chemical specificities of both solute and solvent molecules, such as hydrogen bonding or other association and hydrophobic forces, by yielding the 3D density distributions of solvent, similar to explicit solvent simulations. Moreover, it readily provides, via analytical expressions, all the solvation thermodynamics, including the solvation chemical potential, its energetic and entropic decomposition, and partial molar volume and compressibility. The theory has been successful in analyzing a number of chemical and biological systems in solution [250], including synthetic organic supramolecules (e.g. organic rosette nanotubes) [263] as well as peptides and proteins [174, 264–268]. It constitutes a promising method to contract solvent degrees of freedom in MD simulation. Miyata and Hirata [244] have introduced a coupling of 3D-RISM with MD in a multiple time step (MTS) algorithm which can be formulated in terms of the impulse based RESPA method. It converges the 3D-RISM equations for the solvent correlations at the current snapshot of the solute conformation by using the accelerated iterative solver (modified direct inversion in the iterative subspace [250]), then performs several MD steps, and solves the 3D-RISM equations over again. The MTS approach was necessary to bring down the relatively large computational expenses of solving the 3D-RISM equations. Their implementation achieved stable simulation with the 3D-RISM equations solved at each 5th step of MD at most, which is not sufficient for realistic simulation of macromolecules and biomolecular structures of interest. In this work, we couple the 3D-RISM solvation theory with MD in the Amber molecular dynamics package in an efficient way that includes a number of accelerating schemes. This includes several cutoffs for the interaction potentials and correlation functions, an iterative guess for the 3D-RISM solutions, and an MTS procedure with solvation forces at each MD step which are extrapolated from the previous 3D-RISM evaluations. This coupled method makes modeling of biomolecular structures of practical interest, e.g. proteins with water in inner pockets feasible. As a preliminary illustration, we apply the method to Alanine-dipeptide in ambient water.

10.2 Theory and Implementation

Solvation free energies, and their associated forces, are obtained for the solute from the 3D reference interaction site model (3D-RISM) for molecular solvation, coupled with the 3D version of the Kovalenko-Hirata (3D-KH) closure [250]. 3D-RISM provides the solvent structure in the form of a 3D site distribution function, $g_\gamma^{\mathrm{UV}}(\mathbf{r})$, for each solvent site, γ. With $g_\gamma(\mathbf{r}) \to 1$, the solvent density distribution $\rho_\gamma(\mathbf{r}) = \rho_\gamma g_\gamma(\mathbf{r})$ approaches the solvent bulk density ρ_γ. The 3D-RISM integral equation has the form

$$h_\gamma^{\mathrm{UV}}(\mathbf{r}) = \sum_\alpha \int d\mathbf{r}' c_\alpha^{\mathrm{UV}}(\mathbf{r}-\mathbf{r}')\,\chi_{\alpha\gamma}^{\mathrm{VV}}(r'), \tag{10.1}$$

where superscripts 'U' and 'V' denote the solute and solvent species respectively; $h(\mathbf{r}) = g(\mathbf{r}) - 1$ is the site-site total correlation function; $c_\alpha^{\mathrm{UV}}(\mathbf{r})$ is the 3D direct correlation function for solvent site α having asymptotics of the interaction potential between the solute and solvent site: $c_\alpha^{\mathrm{UV}}(\mathbf{r}) \propto -u_\alpha^{\mathrm{UV}}(\mathbf{r})/(k_B T)$; and $\chi_{\alpha\gamma}^{\mathrm{VV}}(r)$ is the site-site susceptibility of the solvent, given by

$$\chi_{\alpha\gamma}^{\mathrm{VV}}(r) = \omega_{\alpha\gamma}^{VV}(r) + \rho_\alpha h_{\alpha\gamma}^{VV}(r). \tag{10.2}$$

Here, $\omega_{\alpha\gamma}^{VV}(r)$ is the intramolecular correlation function, representing the internal geometry of the solvent molecules while $h_{\alpha\gamma}^{VV}(R)$ is the site-site radial total correlation function of the pure solvent calculated from the dielectrically consistent version of the 1D-RISM theory (DRISM) [269, 270]. Equation (10.1) is complemented with the 3D-KH closure

$$g_\gamma^{UV}(\mathbf{r}) = \begin{cases} \exp\left(d_\gamma^{UV}(\mathbf{r})\right) & \text{for } d_\gamma^{UV}(\mathbf{r}) \le 0 \\ 1 + d_\gamma^{UV}(\mathbf{r}) & \text{for } d_\gamma^{UV}(\mathbf{r}) > 0 \end{cases} \tag{10.3}$$

where

$$d_\gamma^{UV}(\mathbf{r}) = -\frac{u_\gamma^{UV}(\mathbf{r})}{k_B T} + h_\gamma^{UV}(\mathbf{r}) - c_\gamma^{UV}(\mathbf{r}),$$

and $u_\gamma^{UV}(\mathbf{r})$ is the total potential of the solute acting on solvent site γ, given by the sum of the potentials from all the solute interaction sites i located at $\mathbf{R}_i$,

$$u_\gamma^{UV}(\mathbf{r}) = \sum_i u_{i\gamma}^{UV}(|\mathbf{r} - \mathbf{R}_i|).$$

As with other closure approximations [250], the 3D-RISM equation (10.1) with 3D-KH closure (10.3) possesses an exact differential of the free energy, and thus has a closed analytical expression for the excess chemical potential of solvation

$$\Delta\mu = k_B T \sum_\alpha \rho_\alpha \int d\mathbf{r} \left\{ \frac{1}{2} \left(h_\alpha^{UV}(\mathbf{r})\right)^2 \Theta(h_\alpha^{UV}(\mathbf{r})) - c_\alpha^{UV}(\mathbf{r}) - \frac{1}{2} h_\alpha^{UV}(\mathbf{r}) c_\alpha^{UV}(\mathbf{r}) \right\}, \tag{10.4}$$

where $\Theta(x)$ is the Heaviside function, which results in $(h_\alpha(\mathbf{r}))^2$ being applied only in areas of site density depletion. The solvation chemical potential is generally determined by the Kirkwood "charging" formula with the "switching" parameter λ, which has the following form in the case of the interaction site model,

$$\Delta\mu = k_B T \sum_\alpha \rho_\alpha \int_0^1 d\lambda \int d\mathbf{r} g_\alpha^{UV}(\mathbf{r};\lambda) \frac{\partial u_\alpha^{UV}(\mathbf{r})}{\partial\lambda}. \tag{10.5}$$

The expression for the solvent force acting on each atom of the solute is defined as a derivative of the solvation chemical potential with respect to the solute atom coordinates $\mathbf{R}_i$. By differentiating the expression (10.5), the force is obtained in the general form valid that is for any closure approximation to the 3D-RISM equation,

$$\mathbf{f}^{UV}(\mathbf{R}_i) = -\frac{\partial\Delta\mu}{\partial\mathbf{R}_i} = -\sum_\alpha \rho_\alpha \int d\mathbf{r} g_\alpha^{UV}(\mathbf{r}) \frac{\partial u_{i\alpha}^{UV}(\mathbf{r} - \mathbf{R}_i)}{\partial\mathbf{R}_i}. \tag{10.6}$$

The expression (10.6) has also been obtained, by differentiating a closure to the 3D-RISM equation, for the 3D-KH closure [248] as well as for the 3D-HNC closure [244, 248].

10.2.1 Implementation and Optimization of 3D-RISM

Modifications to the SANDER molecular dynamics module of Amber were minor. Other than calling the RISM3D subroutine, the only modifications were to add in calls for memory allocation and file input/output. A single 3D-RISM calculation is roughly three orders of magnitude slower than a single time step for a system solvated with the same solvent model at the same volume and density. This is not unexpected as 3D-RISM performs a complete sampling of the solvent about the solute. To obtain meaningful sampling of solute conformations it is necessary to optimize the 3D-RISM calculation. To achieve this goal two different optimization strategies were employed: (1) the pre- and post-processing of the solute-solvent potentials, long-range asymptotics and forces was accelerated using a cut-off scheme; and (2) direct calculation of the 3D-RISM solvation forces was avoided altogether by interpolating current force based off of atom positions from previous time steps.

3D-RISM forces on individual solute atoms are smoothly varying and depend solely on their position relative to other solute atoms. Thus, with sufficient force and position data collected from previous time steps, it is possible to interpolate 3D-RISM solvation forces without a full 3D-RISM calculation. This is accomplished by performing least squares fitting to a three-dimensional interpolation model for the force vector of each atom. For the purposes of this calculation, we use a simple linear combination of the new positions of the neighboring atoms relative to the position, $\mathbf{R}$, of the i^{th} atom,

$$\mathbf{f}_i^{UV}(\mathbf{R}) = f_i^0 + \sum_{j=1}^{n} \hat{\mathbf{F}}_{ij} \cdot (\mathbf{R} - \mathbf{R}_j), \tag{10.7}$$

where $\mathbf{f}(0)_i$ is the zero-component fitting parameter, and $\hat{\mathbf{F}}_{ij} \equiv F_{ij}^{ab}$ with $(a,b) = x,y,z$ is the tensor of fitting parameters, giving the "response" of i^{th} atom to the presence of neighboring atoms $j = 1,\ldots,n$ (including the self-term F_{ii}^{kl}) at their previous positions $\mathbf{R}_j$. The range of atoms to sum over, n, can be up to the total number of solute atoms, but may be limited to the atoms within a cutoff radius of atom i. This gives a total of $3n+1$ parameters for each component of the force for solute atom. For each atom, the system is translated and rotated to orient the coordinates relative to the atom in question and two of its (preferably bonded) nearest neighbors. While better models can be devised, this is sufficient to show the capability of the method.

To our knowledge, this is a new approach to MTS methods. Miyata and Hirata [244] used impulse MTS [119] where slowly varying forces are only applied at an integer multiple of the base time step. This effectively introduces large, periodic impulses to the dynamics. This has desirable properties, such as energy conservation. However, it is well known that resonance artifacts limit the MTS step size to 5 fs, after which the method becomes catastrophically unstable [96]. Extrapolative MTS applies a constant force over all intermediate time steps. There are no impulses in this method and it does not conserve energy. LN MTS couples extrapolative MTS with Langevin dynamics to produce stable trajectories for MTS time steps up to 10s of 100s of femtoseconds, provided the forces being extrapolated are slow varying on these time scales [96]. Impulse and extrapolative MTS are implemented for 3D-RISM solvation forces and compared to the interpolative MTS method.

10.3 3D-RISM-KH and Molecular Dynamics Setup

All simulations were carried out in a modified version of Amber [5] with the ff99SB force field [271] using free boundary conditions (FBC) with a 14 Å cutoff for Coulomb interactions

or periodic boundary conditions (PBC) with particle-mesh Ewald (PME) summation [128]. SHAKE [272] and a base time step of $\Delta t = 2$ fs was used throughout. The standard SANDER velocity Verlet integrator [119] was used for NVE simulations and the Langevin integrator [256], with a friction coefficient of 5 or 20 ps^{-1}, for NVT simulations. A cutoff of 14 Å was used for Lennard-Jones and 3D-RISM calculations except where otherwise noted.

Alanine-dipeptide was used to quantify the effects of 3D-RISM optimizations on NVE and NVT simulations. A cubic, 32 Å box was used for all simulations except for GB simulations, which used free boundary conditions. 3D-RISM simulations used the SPC/E water model [70] with either a 0.5 Å (64^3 grid points) or 0.25 Å (128^3 grid points) linear grid spacing and will be referred to as ALA:64 and ALA:128, respectively. Explicit solvent simulations used the rigid TIP3P water model [69] and will be referred to as ALA:TIP. GB simulations employed the GBneck model, corresponding to igb=7 [273] in SANDER, and will be referred to as ALA:GB. Due to the small size of alanine-dipeptide, the 14 Å cutoff is equivalent to having no cutoff for GB and 3D-RISM calculations. All simulations started with 100 ps of equilibration at a constant temperature of 300 K. ALA:64 and ALA:128 employed impulse MTS with a 3D-RISM force evaluation every second time step. Extrapolation and interpolation MTS algorithms were tested with full 3D-RISM calculations every $2^n \Delta t$ with $n = 1 \ldots 13$.

To calculate the effect of grid resolution on numerical artifacts and integration of forces, including net force drift, four additional single time step simulations were performed using the SPC/E model. These had linear grid spacings of 0.5, 0.25, 0.125 and 0.0625 Å corresponding to grid sizes of 64^3, 128^3, 256^3 and 512^3, respectively. These are denoted ALA:64d, ALA:128d, ALA:256d and ALA:512d, respectively. In addition, the solution for each grid size was converged to a residual error tolerance of 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5} and 10^{-6}. Since equilibration does not have an impact on these calculations, the default structure for alanine-dipeptide from TLEAP was used. Due to technical issues with our large memory computer, we used the Numerical Recipes FFT [274] rather than FFTW.

Following equilibration, NVE simulations were run for 8 ps each, using the final positions and velocities of the equilibration runs. Energies were recorded every time step except for MTS simulations where the energy was only recorded when a full 3D-RISM calculation took place. The longest MTS time step used for these simulations was $32\Delta t$. The role of the accuracy of the solution 3D-RISM solution on energy conservation was investigated for the STS method only. The same 8 ps simulations were carried out with MDIIS residual tolerances for c_α^{UV} of 10^{-4}, 10^{-5} and 10^{-6} for both ALA:64 and ALA:128.

STS NVE simulations for ALA:64 were also repeated at residual tolerances of 10^{-4} and 10^{-5} but without optimizations to the potential, force, asymptotics calculations or solution extrapolation. That is, no cutoffs, no MTS were utilized and the previous 3D-RISM solution was used as the initial guess for the next solution but was not extrapolated from multiple previous solutions.

NVT simulations were run for 100 ps each, following equilibration, using the final positions and velocities of the equilibration runs. Energies were recorded every time step except for MTS simulations where the energy was only recorded when a full 3D-RISM calculation took place. Solvation forces and atom coordinates were recorded every time step.

10.4 Results

10.4.1 Net Force Drift Error

The net x, y, and z solvation forces should be zero. While it is possible to set these to zero by subtracting the mass weighted total force from each atom, a non-zero net force still indicates

Tolerance	ALA:64d	ALA:128d	ALA:256d	ALA:512d
		(a) Absolute Net Force		
10^{-2}	(1.6,0.80,2.7)	(0.36,1.6,1.8)	(0.78,1.7,1.9)	(1.4,2.8,1.0)
10^{-3}	(1.1,-0.30,1.1)	(-0.31,0.11,0.12)	(0.037,0.067,0.053)	(0.16,0.23,0.094)
10^{-4}	(1.1,-0.34,1.0)	(-0.35,0.026,0.058)	(0.061,0.0048,0.0081)	(0.033,0.028,0.0076)
10^{-5}	(1.1,-0.37,1.0)	(-0.37,-0.0036,0.058)	(0.032,-0.026,-0.0015)	(0.0039,0.0017,0.0019)
10^{-6}	(1.1,-0.37,1.0)	(-0.37,-0.0050,0.056)	(0.032,-0.027,-0.00071)	(0.00076,-0.0012,0.00082)
		(b) Relative Force Error		
10^{-2}	1.6×10^{-1}	1.2×10^{-1}	1.3×10^{-1}	1.2×10^{-1}
10^{-3}	7.3×10^{-2}	1.6×10^{-2}	4.2×10^{-3}	1.3×10^{-2}
10^{-4}	7.2×10^{-2}	1.6×10^{-2}	2.8×10^{-3}	2.0×10^{-3}
10^{-5}	7.1×10^{-2}	1.7×10^{-2}	1.9×10^{-3}	2.1×10^{-4}
10^{-6}	7.1×10^{-2}	1.7×10^{-2}	1.9×10^{-3}	7.5×10^{-5}
		(c) RMS Force Error		
10^{-2}	$7.1 \times 10^{+0}$	$6.3 \times 10^{+0}$	$6.3 \times 10^{+0}$	$8.3 \times 10^{+0}$
10^{-3}	3.4×10^{-1}	1.0×10^{-1}	1.2×10^{-1}	6.2×10^{-2}
10^{-4}	1.8×10^{-1}	7.5×10^{-3}	7.6×10^{-4}	8.7×10^{-4}
10^{-5}	1.8×10^{-1}	7.4×10^{-3}	5.1×10^{-5}	9.2×10^{-6}
10^{-6}	1.8×10^{-1}	7.6×10^{-3}	5.0×10^{-5}	-
		(d) Solvation Free Energy		
10^{-2}	7.5794	7.3873	7.4024	8.4253
10^{-3}	14.5614	14.4441	14.4574	14.3924
10^{-4}	14.6366	14.5097	14.5090	14.5092
10^{-5}	14.6382	14.5123	14.5121	14.5117
10^{-6}	14.6382	14.5125	14.5120	14.5116

Table 10.1: Net force (kcal/mol/Å) drift solvation free energy in 3D-RISM calculations at different resoltuon and accuracies.

the total error in the forces due to the numerical method employed. In the case of 3D-RISM these errors arise as a result of errors in the solution or integration of g^{UV} due to the grid resolution. Both are particularly sensitive to the rapid changes in solvent density close to the solute.This correlation is clear in Table 10.1.

To quantify the drift we use three calculations. The absolute drift is the total force in each direction.

$$\mathrm{E_{absf}} = (\sum f_x, \sum f_y, \sum f_z) \tag{10.8}$$

The relative drift is

$$\mathrm{E_{rel.}}_f = \sqrt{\frac{(\sum f_x)^2 + (\sum f_y)^2 + (\sum f_z)^2}{\sum \mathbf{f}^2}} \tag{10.9}$$

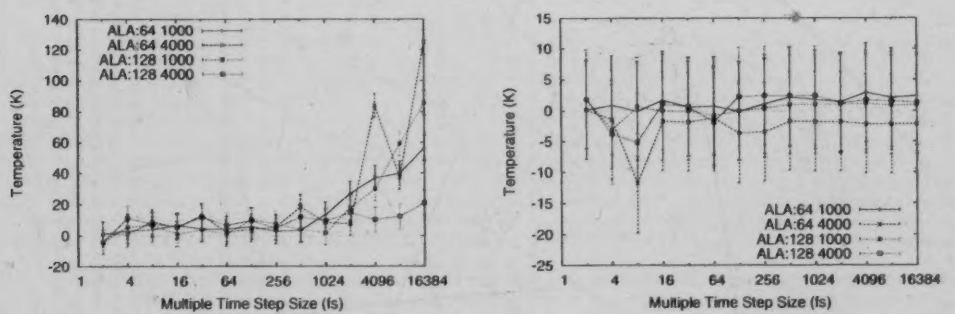
and has the benefit that it can be cheaply calculated at the end of each 3D-RISM calculation. The root-mean-squared error (RMSE) provides an objective comparison to the 'correct' forces, $\tilde{\mathbf{f}}$,

$$\mathrm{RMSE}_f = \sqrt{\frac{\sum \left(\mathbf{f} - \tilde{\mathbf{f}}\right)^2}{N_{\mathrm{sol}}}}. \tag{10.10}$$

Since there is no analytic calculation of the forces available for comparison, we use ALA:512d solved to a 10^{-6} residual error tolerance as our benchmark.

Tolerance	ALA:64	ALA:128	ALA:GB
	Decay rate (kcal/mol/ps)		
	-	-	-0.00637(6)
10^{-4}	-0.4372(9)	-0.2207(6)	-
10^{-5}	-0.0828(6)	-0.0824(6)	-
10^{-6}	-0.0234(6)	-0.0122(5)	-

Table 10.2: Rate of decay for constant energy simulations. Error in the last significant digit is given in brackets.



(a) (b)

Figure 10.1: Error in average temperature for interpolative MTS runs. (a) Langevin friction coefficient of 5 ps^{-1}. (b) Langevin friction coefficient of 20 ps^{-1}. Error bars are block average errors in the mean [119].

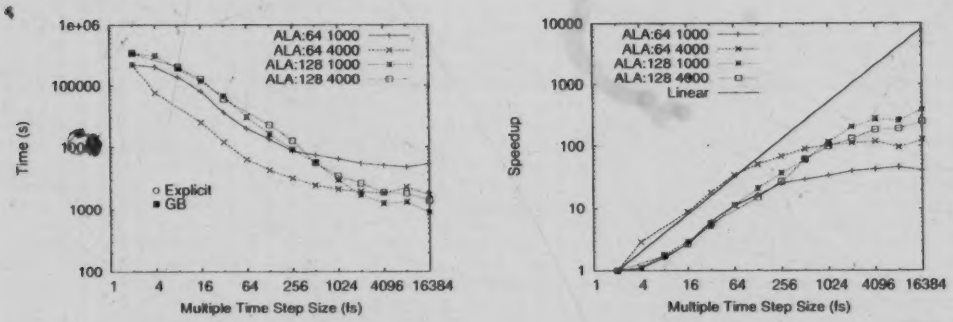
10.4.2 Energy Conservation

The total energy for all constant energy simulations displayed small amplitude oscillations about a linear decay (data not shown). To quantify the linear decay, the equation

$$E_{\text{tot}} = a \cdot t + b \tag{10.11}$$

was fit to each data set with t representing the time in ps and a corresponding to the rate of decay in kcal/mol/ps.

As expected, lower tolerances for the residual of the solution led to better energy conservation but at a higher computational cost. Since energy drifts are system dependent, we must compare 3D-RISM to commonly used methods for the same system. Decay rates for ALA:TIP, ALA:GB, ALA:64 and ALA:128 simulations are given in Table 10.2. While interpolative MTS did show energy conservation comparable to STS up to 10s of ps, all of the simulations gained energy and eventually became unstable (data not shown). The rate of heating was not linear and longer MTS time steps often displayed less heating. Using 4000 instead of 1000 data points did reduce the heating generated. In fact, using a 512 fs 3D-RISM time step showed less absolute change in the energy over an 8 ps run than for ALA:128 with a 10^{-4} residual tolerance with STS. Extrapolative MTS is known not to conserve energy and was not tested for energy conservation as a result.



(a) (b)

Figure 10.2: Speedup resulting from interpolative MTS with Langevin friction coefficient of 20 ps^{-1}. (a) gives the time required for 50 ps of MD. Comparison times for explicit solvent and GB simulations (with an impulse MTS of 4 fs) are labeled on the figure. (b) Speedups relative to STS 3D-RISM with $\Delta t = 2$ fs.

10.4.3 Optimization and MTS Speedup

Speedup, due to the optimization of potential, force and asymptotics calculations, was measured by repeating the NVE simulation of ALA:64 with a residual tolerance of 10^{-4} and 10^{-5}. The unoptimized 3D-RISM calculation required 67% and 27% more time, respectively, to complete an 8 ps simulation.

For constant temperature MD, larger 3D-RISM time steps for both extrapolative and interpolative MTS exhibited greater system heating. Impulse MTS was not tested beyond 4 fs time steps as resonance artifacts exclude larger time steps. As extrapolative MTS also showed considerable heating, these results are not shown. Deviations of the average temperature from the target are reported for interpolative MTS in Figure 10.1. Block averaging is used to calculate errors in the mean and avoid correlation artifacts [119]. Interpolative MTS can produce significant accelerations, bringing 3D-RISM-MD to near the performance of explicit solvent and GB MD (Figure 10.2).

10.5 Discussion

10.5.1 Factors Affecting Numerical Accuracy

In the absence of a heat bath, MD simulates isolated systems. As such, systems simulated with MD should conserve energy and have no net force, within numerical accuracy. 3D-RISM satisfies both of these conditions if the solution is of high enough numerical precision.

The numerical quality of the 3D-RISM solution is controlled by two parameters; (a) the residual error tolerance is the maximum allowable change between two 3D-RISM iterations for convergence and (b) the linear grid spacing is resolution of the grid that the solution is found on. To large extent, these two parameters independently control the conservation of energy and the net force error, respectively.

Table 10.1 gives the error in the net force using a number of different metrics. All three

metrics show that the residual tolerance contributes little to the overall error in the net force until very high grid resolutions are used. This observation is also evident in the solvation free energies calculated. The most important contributor is achieving a minimum resolution of 0.25 Å. Additional iterations to lower the residual provide little improvement to the net force error or the solvation energy calculations.

In practice, the net force is subtracted off as

$$\mathbf{f}_i' = \mathbf{f}_i - \frac{m_i}{M}\mathbf{f}_{\text{net}} \tag{10.12}$$

where m_i is the mass of the i^{th} solute particle and M is the total mass of the solute. However, the error in the net force is also an indicator of inaccuracies in other components not as easily corrected, such as the net torque.

The importance of an accurate solution with a low residual can be observed in Table 10.2. Energy conservation is dominated by the accuracy of the solution, whereas grid resolution has little impact. Here, we can see that a residual tolerance on the order of 10^{-6} give similar energy conservation to that of GB.

Parameters for the grid spacing and residual tolerance must also be chosen with computational practicality and efficiency in mind as well. The computational time required to converge on a solution roughly doubles as the residual tolerance decreases by a factor of 10. I.e. a solution with a residual of 10^{-5} takes twice as long to compute as one with a residual of 10^{-4}. Similarly, both the time and memory requirements of the computation scale as $\mathcal{O}(N)$ where N is the linear number of grid points. Decreasing the grid spacing by a factor of two incurs an eight-fold increase in memory and time requirements.

In light of the computational costs of 3D-RISM calculations and the requirements for performing MD, a reasonable trade off to get reliable results is a grid spacing of 0.25 Å and a residual tolerance of 10^{-5}. However, an acceptable precision of the polarization energy is obtained with a grid spacing of 0.5 Å and a residual tolerance of 10^{-3}, as seen from our preliminary calculations (see also Table 10.1). A detailed account of the polarization energy accuracy for various systems will be given in forthcoming publications.

10.5.2 Speedup

Optimizations for 3D-RISM-MD can be separated into those that use MTS and those that do not. The latter are conservative in that they preserve the overall numerical precision of the method but offer only moderate increases in performance. MTS methods, on the other hand, can provide considerable speedups but sacrifice energy conservation. We analyze the benefits from each of these approaches separately.

Cutoff and solution extrapolation methods were compared against an unoptimized 3D-RISM-MD. The success of the methods varies with the system being simulated. The cutoff methods scale with the number of grid points and solute atoms, just as the unoptimized methods do, as all grid points must still be visited. The major difference is that the number of grid points requiring expensive calculations, involving all of the solute atoms, is considerably reduced using cutoffs. As the grid density and number of solute atoms increases, the cutoff optimizations become more valuable. The solution extrapolation method provides most of the speed gains against the unoptimized code. This is because it works to reduce the number of iterations needed for 3D-RISM convergence. However, as the residual tolerance is lowered, the method becomes less effective. The method cannot propagate solutions at a high accuracy.

All MTS methods exhibited heating for time step sizes beyond 4 fs. However, the heating present in Figure 10.1(a) is effectively absent when the friction coefficient is increased from 5 to 20 ps^{-1}. While it is true that increasing the friction coefficient has a negative impact on

the accuracy of dynamics, the use of a mean-field method, 3D-RISM, means that the observed dynamics are not true dynamics in any case. Our goal is to increase sampling efficiency and using a large friction coefficient is justified in this context.

MTS speedups were measured with all other optimizations included. The method does not increase linearly with the size of the 3D-RISM time step and begins to plateau for multi-picosecond step sizes. The main reason is that conformational correlations are lost for larger time step sizes and the solution extrapolation loses effectiveness. However, accelerations of nearly two orders of magnitude can reasonably be achieved with even the simple model for coordinate based extrapolation used here. It is also of interest to compare theses results to the time required for the equivalent system simulated with explicit solvent or GB. We see in Figure 10.2(a) that the largest MTS time steps for 3D-RISM are comparable to these other methods. Not shown in this figure is the cost of non-bond list updates, which dominate the calculation time, typically consuming over 1200 s. GB suffers the most, with over 99% of calculation time spent doing list updates. However, 3D-RISM accounts for less than half of the run time at time steps this large, accounting, in part, for the non-linear speedup.

10.6 Conclusions

We have presented an efficient coupling of molecular dynamics simulation with the three-dimensional molecular theory of solvation (3D-RISM-KH), contracting the solvent degrees of freedom, and have implemented this multiscale method in the Amber molecular dynamics package.

The 3D-RISM-KH theory uses the first principles of statistical mechanics to provide proper account of molecular specificity of both the solute biomolecule and solvent. This includes such effects as hydrogen bonding both between solvent molecules and between the solute and solvent, as well as hydrophobic hydration and hydrophobic interaction. The 3D-RISM-KH theory readily addresses electrolyte solutions and mixtures of liquids of given composition and thermodynamic conditions. As the solvation theory works in full statistical ensemble, the coupled method yields solvent distributions without statistical noise and gives access to slow processes like hydration of inner spaces and pockets of biomolecules.

The implementation includes several procedures to maximally speed up the calculation: (i) cut-off procedures for the Lennard-Jones and electrostatic potentials and the forces acting on the solute, (ii) cut-offs and approximations for the asymptotics of the 3D site correlation functions of solvent, (iii) an iterative guess for the solution to the 3D-RISM-KH equations at next MD steps by extrapolating the past solutions, (iv) multiple time step (MTS) interpolation of solvation forces between the successive 3D-RISM-KH evaluations of the forces, which are then extrapolated forward at the MD steps until the next 3D-RISM evaluation.

As a preliminary validation, we have applied the method to Alanine-dipeptide in ambient water. Analysis of the accuracy of forces, energy and temperature, including such known artifacts as net force drift, has been performed; factors affecting the accuracy have been quantified and the range of grid resolution and tolerance parameters ensuring reliable results has been outlined. The performance of the coupled method has been characterized, compared to MD with explicit and implicit solvent. This work is a preliminary but significant step toward the full-scale characterization and analysis of the new method, and further improvement of its performance to address slow processes of large biomolecules in solution.

Chapter 11

Summary and Future Work

Protein modelling is a large and diverse field. This thesis has addressed three individual topics relating to protein hydration, microtubule function and improving computational efficiency of molecular modelling. The original motivation was to understand the molecular basis of microtubule function. Modelling tubulin, like any other protein, requires the accurate modelling of solvent, itself a demanding computational task. This naturally leads to the development of optimizing techniques for the accurate and efficient modelling of water.

Myoglobin is, historically, one of the most thoroughly studied proteins, both in general and the specific case of protein hydration. Its neutral net charge at pH 7.0, and well characterized behaviour, make it an ideal subject for the study of hydration and the global, electrostatic properties of both water and protein. Polarization of the electric fields of both protein and water increase as hydration increases. When full hydration occurs at 0.35 h, the polarization of the water plateaus, as do the fluctuations and deviations of protein atoms. Myoglobin's influence over the water is most clearly observed in the first hydration layer around the protein. The orientation of waters within this first layer are strongly correlated with the direction of the electric field of the protein, but this quickly decays as distance from the protein increases. In fact, these surface waters can be significantly decoupled from the surrounding water when very close to the protein. In general, both the dependence on the level of hydration and the proximity to the protein are reduced as the temperature is decreased. This is likely due to the formation of an amorphous glass at low temperatures.

Tubulin is the constituent protein in microtubules (MTs) and the basis for their stability and structure. Tubulin is also a challenging molecule to simulate as it is a large protein that is slow to relax. The poor quality of the available crystal structures compounds this problem. However, we are interested in tubulin as a part of MTs; thus, interactions between multiple dimers are the ultimate concern. Approaching this problem through the use of truncated systems and periodic boundary conditions, we have investigated three key reaction coordinates and calculated the potential of mean force (PMF) along them.

All-atom modelling of microtubule stability has, so far, been limited to protofilament offset and lattice type [173, 174]. Even here the methods have used rigid structures for tubulin. We have used a fully flexible model, with only centre of mass and orientational restrains to maintain an MT geometry. These simulations demonstrate the importance of the M-loop in contributing to MT flexibility and stability. The flexibility of the M-loop allows it to deform as the protofilaments change their longitudinal offset and maintain contact the with H1-S2 loop on the adjacent protofilament. This lends flexibility and stability to MTs under shear force. It also suggests a role for MT stabilizing agents, such as Taxol and epothilone. These drugs may stabilize the M-loop to encourage interaction with the adjacent dimer without impairing

flexibility along the longitudinal axis. However, this same deformation in the M-loop prevents the convergence of our calculations. The stretching of this loop leads to a strong hysteresis that can only over come with the adoption of a multidimensional reaction coordinate.

While the lack of convergence in these simulations is unfortunate, its cause, the flexibility of the M-loop, offers significant insights to the mechanical properties of MTs and suggests new avenues of research. A 2D reaction coordinate could be used to fully explore the PMF of protofilament offset but a more direct investigation of the M-loop and the effect of MT stabilizing drugs is more immediately of interest and computationally accessible. Using the same reaction coordinates (longitudinal offset and lateral separation but constrained to only a few Ångströms of longitudinal offset) with and without Taxol or epothilone the effect on the M-loop and any resulting change in the PMF can be readily observed. This can be compared against experiment and, possibly, explain the role of these ligands on MT rigidity. As computers get faster and algorithms more efficient, this could be revisited for the entire protofilament offset reaction coordinate.

The switch of MTs from stable to unstable has been the focus of considerable research. A consensus view has been established of a conformational change in tubulin from a straight (or slightly bent) conformer when GTP is bound to a bent conformer when GDP is present [15, 16, 31, 176]. This agrees well with measurements that have shown that polymerizing MTs can generate force [275] and coarse-grained models that incorporate this conformational change can explain it [276, 277]. Our calculations, while only partially converged, suggest an alternative mechanism. Rather than a large conformational change, where dimers adopt a distorted, bent conformation, the attraction at the inter-dimer interface simply weakens, allowing the protofilament to become more flexible and explore more space. This softened PMF does not produce any depolymerizing force itself. However, in the context of a MT, an entropic force is produced as a depolymerized MT is more entropically favourable than an assembled one. As the individual dimers are still in a lattice, they cannot simply fall apart. Rather, the protofilaments fall away. As they cannot fall into the MT the entropic force is directed outward, producing the characteristic 'ram's horns'. Thus, MT depolymerization may be explained through a subtle shift in an attractive interaction, tipping the balance between stability and instability.

It is also of interest that a subtle change in conformation does occur. The S3-H3 loop, or switch-II loop, is slightly extended in the presence of GTP. This extension strengthens interactions with the adjacent dimer and stabilizes the interface. This loop has long been thought to be involved in conformational change due to its similarity to the switch-II loop in classical GTPases [278]. The conformational change, however, is much more subtle than originally thought.

Convergence of these simulations is simply a matter of time. Once converged, the simulations will provide deep insights to effects of GTP hydrolysis and MT stability. However, the 1D reaction coordinate provides only partial information. A 2D reaction coordinate, with radial and tangential displacements in the MT coordinate system, would give a more complete picture and explore the possibility of bend angles perpendicular to our 1D reaction coordinate. Combined with a similar investigation of the M-loop, as discuss previously, a simple model to explain the elastic properties of MTs can be constructed.

Sequence differences between isotypes are known to modify the physical properties of MTs. As isotypes are difficult to work with, little is known about the details of the differences in the properties or how specific changes contribute to these changes. As it is known that the most sequence variation between isotypes is in the carboxy-terminal tail (CTT) region of tubulin we have focused on isotype differences for this region of the protein through rigorous conformational sampling for all nine human isotypes of β-tubulin. Using replica exchange and principal component analysis methods we were able to characterize the physical properties of

the isotypes, including flexibility, average end-to-end distance, secondary structure and ρ_{sc}. A range of properties was observed with β-III and β-VI being the most flexible with the longest end-to-end distance while β-VII was significantly shorter and stiffer than the other fragments. Binding motifs common to several isotypes were also identified for MAP2 and Casein Kinase-2 based off of sequence and conformation alone.

While this study characterizes the basic physical properties of the CTTs, isotype differences between the tails in the presence of tubulin and microtubule associated proteins (MAPs), such as kinesin, is not known. Due to the computational burden created from simulating these large molecules, the methodology employed must be modified. In particular, to use replica exchange MD, the number of degrees of freedom must be reduced. This can be accomplished by using an implicit solvent or, ideally, 3D-RISM. As the bulk of tubulin, and any bound MAP, is effectively rigid, we are only interested in the CTT. We can ignore these extraneous degrees of freedom by using a method such as partial replica exchange MD [148]. In this case the CTT would be coupled to a different heat bath than the rest of the system, which would maintain the target temperature in all replicas. The number of degrees of freedom are then significantly reduced and the CTT in the context of these physiologically meaningful systems can be explored.

The computational resources required for simulations such as these are a significant impediment to performing large scale studies. As water typically accounts for 95% of the atoms in such a system, an obvious approach is to eliminate the explicit water in our simulations. Methods such as Poisson-Boltzmann and generalized-Born have been used to accomplish this but, as they are models of a featureless, continuum dielectric, they introduce artifacts into simulations. The 3D reference interaction site model (3D-RISM) is an alternate approach, based on the Ornstein-Zernike equation of molecular solvation, that accounts for the detailed, microscopic structure of solvents such as water. We have implemented this model in the molecular dynamics package Amber and carried out optimizations to reduce the computational requirements of 3D-RISM by over two orders of magnitude. These optimizations include simple cutoffs (that preserve long range behaviour), propagators of the solution to the 3D-RISM integral equations and multiple time step methods. The latter uses a new, interpolative scheme that is uniquely suited to 3D-RISM and is responsible for the large time step and low computational load of 3D-RISM. When using extremely large time steps (e.g. > 8 ps), the method requires CPU time comparable to explicit solvent and removes nearly 95% of the degrees of freedom. This makes the 3D-RISM/MD combination particularly suited for replica exchange MD.

3D-RISM coupled with MD is in its infancy and displays enormous potential that must be developed. Several immediate issues must be addressed. The method must be parallelized to work on distributed memory system. This is essential for speed and to distribute the memory requirements to multiple machines, allowing larger systems to be simulated without specialized hardware. Aiding the simulation of large systems will be the development of a 'minimal' solvation box, numerically calculating the first solvation shell while accounting for the long-range solvation structure analytically. This will reduce memory and computation time for 3D-RISM solutions several fold. An improved model for the interpolative force calculation is also required. Ideally, it should be physically based and conserve energy.

As the 3D-RISM/MD method matures, it should be coupled with methods that are best suited to its strengths. A principal advantage of statistical-mechanical, molecular theory of solvation (3D-RISM) is that it accounts for the solvent effect from the whole statistical-mechanical ensemble, both over space and time. The 3D-RISM theory effectively distributes the solvent (including co-solvent and possible guest species, such as drug molecules) around the protein, including pockets and cavities. Achieving this directly with molecular simulations would require several opening (partially 'unfolding') and closing ('refolding') events to allow solvent molecules to pass in and out. The use of 3D-RISM allows one to bypass this process, and

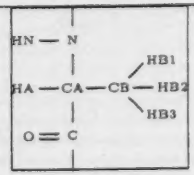
get the solvent distribution and solvation free energy for individual conformations, including all the pockets and inner spaces. MD can then be concentrated on the conformational search for the folded protein around the pool of target conformations, bypassing rarely sampled intermediate or unfavourable transition states, with the solvent structure and thermodynamic contribution provided by 3D-RISM on the fly for the target conformations. This enormously extends feasibility of simulating processes of docking, protein interactions and functions. 3D-RISM can be further exploited through use in replica exchange MD (REMD), where reduction in the number of replicas required and the removal of friction aids sampling and can allow the method to be somewhat slower than explicit solvation. The method can also be adapted to use free or periodic boundary conditions in each dimension independently, making it suitable for membrane simulations.

All of these properties can be put to practical use in the simulation of tubulin. While tubulin does not have any internal, solvated cavities the inter- and intra-dimer interfaces. Solvating and desolvating these interfaces introduces significant hysteresis into simulations calculating binding free energies of binding. 3D-RISM eliminates this problem entirely. Simulations of tubulin protofilaments also benefit from the use of periodic boundary conditions. By applying these condition only along the protofilament axis, it is possible to avoid artifacts due to periodicity in the other dimensions. Finally, using 3D-RISM in REMD will drastically reduce the number of replicas required to simulate it. Further extending this with a methodology like partial REMD [148] will allow the local refinement and study of C-terminal tails in the physiological context of tubulin and microtubule associate proteins. By continuing to develop 3D-RISM/MD, large scale, accurate simulations of tubulin and microtubules will become possible.

Appendix A

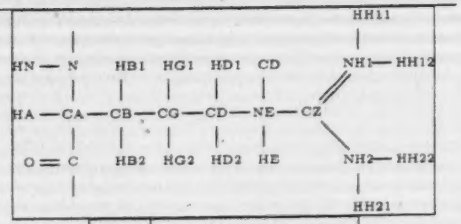
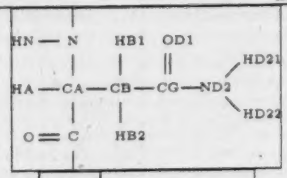
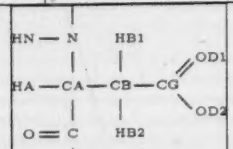
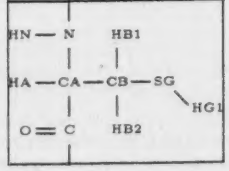
Amino Acids

Table A.1: Amino acid abbreviations and properties[6, 7, 279]. Chemical structures show the IUPAC names for the atoms for the amino acids that are not N or C-terminii[120]. Since histidine is likely to be either singly or doubly protonated near pH 7.0 the two candidates for de-protonation are highlighted in red. Either, but not both, may be removed. Histidine is given a net charge of $1e$ at pH 7.0 since it is a more probable state, though not as decisively as for other residues.

Amino Acid	Three-letter Abbreviation	One-letter Symbol	Group	Typical Charge at pH 7.0 (e)	Structure
Alanine	Ala	A	nonpolar	0	

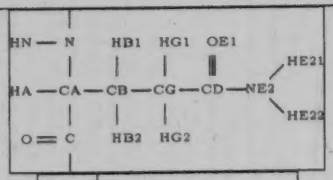
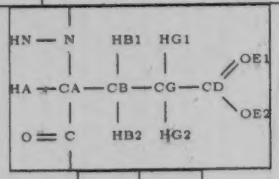
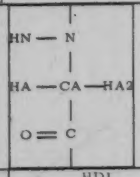
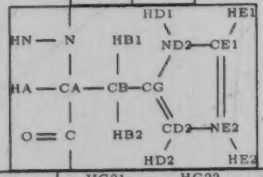
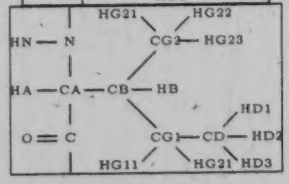
continued on next page

continued from previous page

Amino Acid	Three-letter Abbreviation	One-letter Symbol	Group	Typical Charge at pH 7.0 (e)	Structure
Arginine	Arg	R	basic	1	
Asparagine	Asn	N	polar	0	
Aspartic Acid	Asp	D	acidic	-1	
Cysteine	Cys	C	nonpolar /special	0	

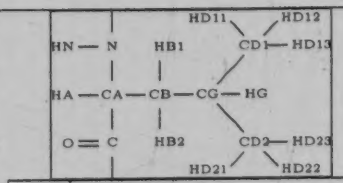
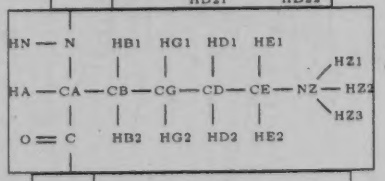
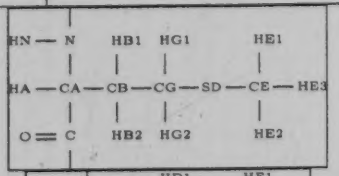
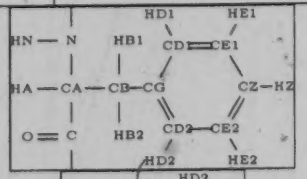
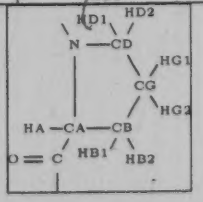
continued on next page

continued from previous page

Amino Acid	Three-letter Abbreviation	One-letter Symbol	Group	Typical Charge at pH 7.0 (*e*)	Structure
Glutamine	Gln	Q	polar	0	
Glutamic Acid	Glu	E	acidic	-1	
Glycine	Gly	G	nonpolar /special	0	
Histidine	His	H	basic	0	
Isoleucine	Ile	I	nonpolar	0	

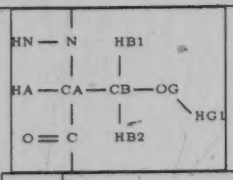
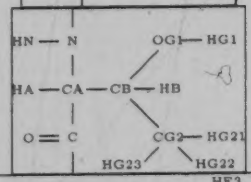
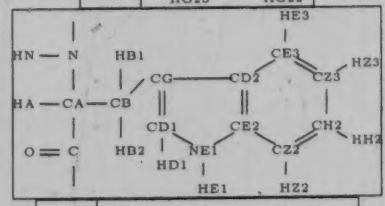
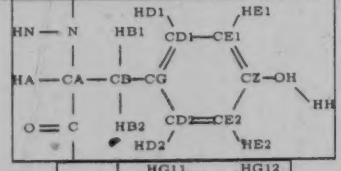
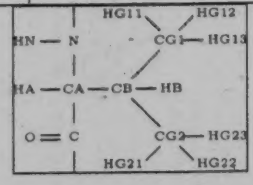
continued on next page

continued from previous page

Amino Acid	Three-letter Abbreviation	One-letter Symbol	Group	Typical Charge at pH 7.0 (ϵ)	Structure
Leucine	Leu	L	nonpolar	0	
Lysine	Lys	K	basic	1	
Methionine	Met	M	nonpolar	0	
Phenylalanine	Phe	F	nonpolar	0	
Proline	Pro	P	nonpolar /special	0	

continued on next page

continued from previous page

Amino Acid	Three-letter Abbreviation	One-letter Symbol	Group	Typical Charge at pH 7.0 (e)	Structure
Serine	Ser	S	polar	0	
Threonine	Thr	T	polar	0	
Tryptophan	Trp	W	nonpolar	0	
Tyrosine	Tyr	Y	nonpolar	0	
Valine	Val	V	nonpolar	0	

Appendix B

Protein Net charge

In the reversible reaction

$$HA \rightleftharpoons H^+ + A^-, \tag{B.1}$$

where A^- is a deprotonated acid, HA is a protonated acid and H^+ is a proton, we can define an equilibrium constant K_a as [279]

$$K_a = \frac{[H^+][A^-]}{[HA]}. \tag{B.2}$$

Variables in square braces refer to the concentration of those substances.

We are interested in the total amount of the acid species in solution, rather than the amount of protonated acid in solution. This is easily obtained from

$$[A_T] = [HA] + [A^-] \tag{B.3}$$

$$[A^-] = [A_T] - [HA] \tag{B.4}$$

where $[A_T]$ is the total concentration of the acid species. Substituting this into Equation (B.2) we obtain

$$K_a = \frac{[H^+][A_T]}{[HA]} - [H^+]. \tag{B.5}$$

We then solve for the number of protonated acids:

$$[HA] = \frac{[H^+][A_T]}{K_a + [H^+]}. \tag{B.6}$$

Since we know the number of each residue in the protein, and not the concentration, we multiply both sides by the volume

$$N_{HA} = \frac{[H^+]N_{A_T}}{K + [H^+]}. \tag{B.7}$$

It is now convenient to work in pH and pK_a where 'p' denotes '$-\log_{10}$' and we rewrite Equation (B.7) as

$$N_{HA} = \frac{10^{-pH}N_{A_T}}{10^{-pK_a} + 10^{-pH}}. \tag{B.8}$$

Table B.1: pK_a values for ionization of seven Amino Acids at 25° C [279].

Amino Acid	pK_a
Aspartic acid	3.86
Glutamic acid	4.25
Histidine	6.0
Cysteine	8.33
Tyrosine	10.07
Lysine	10.53
Arginine	12.48

Seven of the 20 standard amino acids and both the N terminus (-NH_3) and C terminus (-COOH) are proton donors. The probability of the proton being donated depends on the pH. The pH at which the proton is equaly likely to be attached to the protein or be in solution is denoted as pK_a. The pK_as for the proton donors can be found in Table B.

Of these 9 proton donors lysine, arginine, histidine and the N-terminus are positively charged when fully protonated while the rest are neutral. This means that when counting the total charge of the protein we must count the positively charged protonated residues separately from the neutral prototated residues. For lysine, arginine, histidine and the N-terminii we record the total number protonated residues. For aspartic acid, glutamic acid, cysteine, tyrosine and the C-terminii we count the number of protonated residues less the total number of these residues. Thus, we have for a general protein

$$\begin{aligned}\text{Total Charge} =& N_{lys+H} + N_{arg+H} + N_{his+H} + N_{NH_3^+} \\ &+ (N_{asp+H} - N_{asp}) + (N_{glu+H} - N_{glu}) \\ &+ (N_{cys+H} - N_{cys}) + (N_{tyr+H} - N_{tyr}) \\ &+ (N_{COOH} - N_{COO}).\end{aligned} \tag{B.9}$$

Using Equation (B.8) this becomes

$$\begin{aligned}\text{Total Charge} =& \frac{10^{-pH} N_{lys}}{10^{-pK_{a,lys}} + 10^{-pH}} + \frac{10^{-pH} N_{arg}}{10^{-pK_{a,arg}} + 10^{-pH}} \\ &+ \frac{10^{-pH} N_{his}}{10^{-pK_{a,his}} + 10^{-pH}} + \frac{10^{-pH} N_{NH_{2/3}}}{10^{-pK_{a,NH_{2/3}}} + 10^{-pH}} \\ &+ \frac{10^{-pH} N_{asp}}{10^{-pK_{a,asp}} + 10^{-pH}} + \frac{10^{-pH} N_{glu}}{10^{-pK_{a,glu}} + 10^{-pH}} \\ &+ \frac{10^{-pH} N_{cys}}{10^{-pK_{a,cys}} + 10^{-pH}} + \frac{10^{-pH} N_{tyr}}{10^{-pK_{a,tyr}} + 10^{-pH}} \\ &+ \frac{10^{-pH} N_{COO}}{10^{-pK_{a,COO}} + 10^{-pH}} \\ &- (N_{asp} + N_{glu} + N_{tyr} + N_{cys} + N_{COO}).\end{aligned} \tag{B.10}$$

This generalizes to

$$\begin{aligned}\text{Total Charge} =& \sum_{\text{Proton Donors}} \frac{10^{-pH} N_{residue}}{10^{-pK_{a,residue}} + 10^{-pH}} \\ &+ \sum_{\text{Neutral Proton Donors}} N_{residue} Q_{residue}\end{aligned} \tag{B.11}$$

where Q is the unprotonated charge of the residue.

While this method does work well in general, particularly for small proteins and peptides, the limiting assumption is that the pK_a for each residue remains the same, regardless of environment. In practice, the local environment, including other amino acids in the protein, causes the pK_as of individual amino acids to shift. There has been considerable work on predicting these shifts using molecular modelling techniques. For example, see Li *et al.* [184, 280] and Mongan *et al.* [281].

Appendix C

CHARMM System of Units

For all of CHARMM's, Amber's and NAMD's internal calculations the AMKA (Ångstroms, Kilocalories/Mole, Atomic mass units) system of units is used[98, 124, 125]. GROMACS uses the same system with kJ instead of kcal [146]. Distances are measured in Ångstroms, energy in kilocalories/mole, masses in atomic mass units and charge in units of elementary charge. A unit of time is then calculated in this system to be 4.888821E-14 seconds.

Using $N_A = 6.02214 \times 10^{23}$/mol, $e = 1.60218 \times 10^{-19}$C, 1J = 2.3901^{-4}kcal, 1m = 10^{10}Å, $k_B = 1.38066 \times 10^{-23}$J/K [282] and $\epsilon_0 = 8.8542 \times 10^{-12}$F/m[283] in AMKA units we have

$$\epsilon_0 = 2.3964 \times 10^{-4} \frac{e^2}{\text{kcal/mol Å}} \tag{C.1}$$

and

$$k_b = 1.9873 \times 10^{-3} \frac{\text{kcal}}{\text{K·mol}}. \tag{C.2}$$

Bibliography

[1] D. Beglov and B. Roux. Numerical solution of the hypernetted chain equation for a solute of arbitrary geometry in three dimensions. *J. Chem. Phys.*, **103**(1), 360–364, 1995.

[2] D. Beglov and B. Roux. An integral equation to describe the solvation of polar molecules in liquid water. *J. Phys. Chem. B*, **101**(39), 7821–7826, 1997.

[3] A. Kovalenko and F. Hirata. Three-dimensional density profiles of water in contact with a solute of arbitrary shape: A RISM approach. *Chem. Phys. Lett.*, **290**(1-3), 237–244, 1998.

[4] A. Kovalenko and F. Hirata. Self-consistent description of a metal–water interface by the kohn–sham density functional theory and the three-dimensional reference interaction site model. *J. Chem. Phys.*, **110**(20), 10095–10112, 1999.

[5] D. Case, T. Cheatham, T. Darden, H. Gohlke, R. Luo, K. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. Woods. The Amber biomolecular simulation programs. *J. Comput. Chem.*, **26**, 1668–1688, 2005.

[6] H. Lodish, A. Berk, S. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, *Molecular Cell Biology*, W. H. Freeman and Co., fourth ed., 2000.

[7] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. Watson, *Molecular Biology of the Cell*, Garland Publishing Inc., fourth ed., 2002.

[8] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne. The protein data bank. *Nucleic Acids Res.*, **28**, 235–242, 2000. http://www.pdb.org/.

[9] S. Parkin and H. Rupp, B.and Hope. Structure of bovine pancreatic trypsin inhibitor at 125 k: Definition of carboxyl-terminal residues gly57 and ala58. *Acta Crystallogr. D*, **52**, 18, 1996.

[10] R. Laskowski, G. Hutchinson, A. Michie, A. Wallace, M. Jones, A. Martin, N. Luscombe, D. Milburn, A. Kasuya, J. Bouquiere, and J. Thornton, Pdbsum.

[11] W. Humphrey, A. Dalke, and K. Schulten. Vmd - visual molecular dynamics. *J. Mol. Graph.*, **14**, 33–38, 1996. http://www.ks.uiuc.edu/Research/vmd/.

[12] M. Paoli, R. Liddington, J. Tame, A. Wilkinson, and G. Dodson. Crystal structure of t state haemoglobin with oxygen bound at all four haems. *J. Mol. Biol.*, **256**, 775, 1996. PDB ID: 1GZX.

[13] E. Nogales, S. Wolf, and K. Dowling. Structure of the $\alpha\beta$-tubulin dimer by electron crystallography. *Nature*, **391**, 199–203, 1998. PDB ID: 1TUB.

[14] J. Löwe, H. Li, K. Dowling, and E. Nogales. Refined structure of $\alpha\beta$-tubulin at 3.5 å resolution. *J. Mol. Biol.*, **313**, 1045–1057, 2001. PDB ID: 1JFF.

[15] R. B. G. Ravelli, B. Gigant, P. A. Curmi, I. Jourdain, S. Lachkar, A. Sobel, and M. Knossow. Insight into tubulin regulation from a complex with colchicine and a stathmin-like domain. *Nature*, **428**(6979), 198–202, 2004.

[16] B. Gigant, C. Wang, R. Ravelli, F. Roussi, M. Steinmetz, P. Curmi, A. Sobel, and M. Knossow. Structural basis for the regulation of tubulin by vinblastine. *Nature*, **435**, 519–522, 2005.

[17] H. Aldaz, L. M. Rice, T. Stearns, and D. A. Agard. Insights into microtubule nucleation from the crystal structure of human gamma-tubulin. *Nature*, **435**(7041), 523–527, 2005.

[18] J. H. Nettles, H. Li, B. Cornett, J. M. Krahn, J. P. Snyder, and K. H. Downing. The binding mode of epothilone A on alpha,beta-tubulin by electron crystallography. *Science*, **305**(5685), 866–869, 2004. PDB ID: 1TVK.

[19] M. Kikkawa, Y. Okada, and N. Hirokawa. 15 A resolution model of the monomeric kinesin motor, KIF1A. *Cell*, **100**(2), 241–252, 2000.

[20] A. Roll-Mecak and R. Vale. Structural basis of microtubule severing by the hereditary spastic paraplegia protein spastin. *Nature*, **451**, 363–367, 2008.

[21] J. Duan and M. Gorovsky. Both carboxy-terminal tails of alpha- and beta-tubulin are essential, but either one will suffice. *Curr. Biol.*, **12**, 313–316, 2002.

[22] H. Li, D. J. DeRosier, W. V. Nicholson, E. Nogales, and K. H. Downing. Microtubule structure at 8*Å* resolution. *Structure*, **10**, 1317–1328, 2002.

[23] H. Detrich, S. Parker, R. Williams, E. Nogales, and K. Downing. Cold adaptation of microtubule assembly and dynamics. Structural interpretation of primary sequence changes present in the alpha- and beta-tubulins of Antarctic fishes. *J. Biol. Chem.*, **275**, 37038–37047, 2000.

[24] C. Modig, M. Wallin, and P. Olsson. Expression of cold-adapted beta-tubulins confer cold-tolerance to human cellular microtubules. *Biochem. Biophys. Res. Commun.*, **269**, 787–791, 2000.

[25] J. Correia, A. Beth, and J. Williams, RC. Tubulin exchanges divalent cations at both guanine nucleotide-binding sites. *J. Biol. Chem.*, **263**(22), 10681–10686, 1988.

[26] D. Chrtien and S. Fuller. Microtubules switch occasionally into unfavorable configurations during elongation. *J. Mol. Biol.*, **298**, 663–676, 2000.

[27] M. Kikkawa, T. Ishikawa, T. Nakata, T. Wakabayashi, and N. Hirokawa. Direct visualization of the microtubule lattice seam both in vitro and in vivo. *J. Cell Biol.*, **127**, 1965–1971, 1994.

[28] F. Metoz, I. Arnal, and R. Wade. Tomography without tilt: three-dimensional imaging of microtubule/motor complexes. *J. Struct. Biol.*, **118**, 159–168, 1997.

[29] E. Unger, K. Böhm, and W. Vater. Structural diversity and dynamics of microtubules and polymorphic tubulin assemblies. *Electron Microsc. Rev.*, **3**, 355–395, 1990.

[30] K. Böhm, W. Vater, P. Steinmetzer, and E. Unger. Effect of sodium chloride on the structure of tubulin assemblies. *Acta Histochem. Suppl.*, **39**, 365–371, 1990.

[31] H. Wang and E. Nogales. Nucleotide-dependent bending flexibility of tubulin regulates microtubule assembly. *Nature*, **435**, 911–915, 2005.

[32] H. Erickson. γ-tubulin nucleation: template or protofilament? *Nat. Cell Biol.*, **2**, E93–E96, 2000.

[33] H. Erickson and D. Stoffler. Protofilaments and rings, two conformations of the tubulin family. *J. Cell Biol.*, **135**, 5–8, 1996.

[34] B. Raynaud-Messina and A. Merdes. Gamma-tubulin complexes and microtubule organization. *Curr. Opin. Cell Biol.*, **19**, 24–30, 2007.

[35] M. Moritz, M. B. Braunfeld, V. Guenebaut, J. Heuser, and D. A. Agard. Structure of the gamma-tubulin ring complex: a template for microtubule nucleation. *Nat. Cell Biol.*, **2**(6), 365–370, 2000.

[36] D. Morgan, *Microtubule Structure and Behavior*, New Science Press Ltd, 2007.

[37] R. Heald and E. Nogales. Microtubule dynamics. *J. Cell. Sci.*, **115**, 3–4, 2002.

[38] V. Sudakin and T. Yen. Targeting mitosis for anti-cancer therapy. *BioDrugs*, **21**, 225–233, 2007.

[39] J. Howard and A. Hyman. Microtubule polymerases and depolymerases. *Curr. Opin. Cell Biol.*, **19**, 31–35, 2007.

[40] E. Niel and J. Scherrmann. Colchicine today. *Joint Bone Spine*, **73**, 672–678, 2006.

[41] M. Jordan and L. Wilson. Microtubules as a target for anitcancer drugs. *Nat. Rev. Cancer*, **4**, 253–265, 2004.

[42] J. T. Huzil, R. F. Luduena, and J. Tuszynski. Comparative modelling of human β-tubulin isotypes and implications for drug binding. *Nanotechnology*, **17**(4), S90–S100, 2006.

[43] E. Carpenter, J. Huzil, R. Luduea, and J. Tuszynski. Homology modeling of tubulin: influence predictions for microtubule's biophysical properties. *Eur. Biophys. J.*, **36**, 35–43, 2006.

[44] T. Luchko, J. Huzil, M. Stepanova, and J. Tuszynski. Conformational analysis of the carboxy-terminal tails of human beta-tubulin isotypes. *Biophys. J.*, **94**, 1971–1982, 2008.

[45] M. Sugiura, R. Maccioni, J. Cann, E. York, J. Stewart, and G. Kotovych. A proton magnetic resonance and a circular dichroism study of the solvent dependent conformation of the synthetic tubulin fragment Ac tubulin, alpha (430-441) amide and its interaction with substance-P. *J. Biomol. Struct. Dyn.*, **4**, 1105–1117, 1987.

[46] A. Otter and G. Kotovych. The solution conformation of the synthetic tubulin fragment Ac-tubulin-α(430-441)-amide based on two-dimensional ROESY experiments. *Can. J. Chem.*, **66**, 1988.

[47] R. F. Ludueña and A. Banerjee In Fojo [284], chapter 6.

[48] R. F. Ludueña and A. Banerjee In Fojo [284], chapter 5.

[49] S. Westermann and K. Weber. Post-translational modifications regulate microtubule function. *Nat. Rev. Mol. Cell Biol.*, **4**, 938–947, 2003.

[50] T. Sarkar, T. Manna, S. Bhattacharyya, P. Mahapatra, A. Poddar, S. Roy, J. Pena, R. Solana, R. Tarazona, and B. Bhattacharyya. Role of the carboxy-termini of tubulin on its chaperone-like activity. *Proteins: Struct. Func. Genet.*, **44**, 262–269, 2001.

[51] M. Baumann, T. Wisniewski, E. Levy, G. Plant, and J. Ghiso. C-terminal fragments of alpha- and beta-tubulin form amyloid fibrils in vitro and associate with amyloid deposits of familial cerebral amyloid angiopathy, British type. *Biochem. Biophys. Res. Commun.*, **219**, 238–242, 1996.

[52] P. Kuhn, M. Knapp, S. Soltis, G. Ganshaw, M. Thoene, and R. Bott. The 0.78 å structure of a serine protease: Bacillus lentus subtilisin. *Biochemistry*, **37**(39), 13446–13452, 1998.

[53] T. Prangé, M. Schiltz, L. Pernot, N. Colloc'h, S. Longhi, W. Bourguet, and R. Fourme. Exploring hydrophobic sites in proteins with xenon or krypton. *Proteins: Struct. Func. Genet.*, **30**, 61–73, 1998.

[54] L. Knipling, J. Hwang, and J. Wolff. Preparation and properties of pure tubulin S. *Cell Motil. Cytoskeleton*, **43**, 63–71, 1999.

[55] S. Rai and J. Wolff. The C terminus of beta-tubulin regulates vinblastine-induced tubulin polymerization. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 4253–4257, 1998.

[56] D. Sackett, B. Bhattacharyya, and J. Wolff. Tubulin subunit carboxyl termini determine polymerization efficiency. *J. Biol. Chem.*, **260**, 43–45, 1985.

[57] A. Marx, J. Mller, E. Mandelkow, A. Hoenger, and E. Mandelkow. Interaction of kinesin motors, microtubules, and MAPs. *J. Muscle Res. Cell. Motil.*, **27**, 125–137, 2005.

[58] S. Lakämper and E. Meyhöfer. The E-hook of tubulin interacts with kinesin's head to increase processivity and speed. *Biophys. J.*, **89**, 3223–3234, 2005.

[59] G. Skiniotis, J. Cochran, J. Mller, E. Mandelkow, S. Gilbert, and A. Hoenger. Modulation of kinesin binding by the C-termini of tubulin. *EMBO J.*, **23**, 989–999, 2004.

[60] M. Kikkawa, E. P. Sablin, Y. Okada, H. Yajima, R. J. Fletterick, and N. Hirokawa. Switch-based mechanism of kinesin motors. *Nature*, **411**(6836), 439–445, 2001.

[61] M. Kikkawa and N. Hirokawa. High-resolution cryo-EM maps show the nucleotide binding pocket of KIF1A in open and closed conformations. *EMBO J.*, **25**, 4187–4194, 2006.

[62] K. Mukhopadhyay, P. Parrack, and B. Bhattacharyya. The carboxy terminus of the alpha subunit of tubulin regulates its interaction with colchicine. *Biochemistry*, **29**, 6845–6850, 1990.

[63] J. Finney. Water? What's so special about it? *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, **359**, 1145–1163, 2004.

[64] J. L. Finney. The water molecule and its interactions: the interaction between theory, modelling, and experiment. *J. Mol. Liq.*, **90**(1-3), 303–312, 2001.

[65] B. Guillot. A reappraisal of what we have learnt during three decades of computer simulations on water. *J. Mol. Liq.*, **101**(1-3), 219–260, 2002.

[66] ed. D. P. Lide, *CRC Handbook of Chemistry and Physics*, CRC PRESS, 88 ed., 2008.

[67] W. S. Benedict, N. Gailar, and E. K. Plyler. Rotation-vibration spectra of deuterated water vapor. *J. Chem. Phys.*, **24**(6), 1139–1165, 1956.

[68] Y. Y. Efimov. Correlations between bond lengths and stretching frequencies in the O-D...O hydrogen bridge and their consequences. *Russian Chem. B.*, **52**(1), 261–264, 2003.

[69] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, **79**(2), 926–935, 1983.

[70] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma. The missing term in effective pair potentials. *J. Phys. Chem.*, **91**(24), 6269–6271, 1987.

[71] A. K. Soper and M. G. Phillips. A new determination of the structure of water at 25-degrees-c. *Chem. Phys.*, **107**(1), 47–60, 1986.

[72] C. von Grotthuss. Sur la décomposition de l'eau et des corps qu'elle tient en dissolution à l'aide de l'électricité galvanique. *Ann. Chim.*, pp. 54–73, 1806.

[73] R. Baldwin. Energetics of protein folding. *J. Mol. Biol.*, **371**, 283–301, 2007.

[74] Y. Levy and J. N. Onuchic. Water mediation in protein folding and molecular recognition. *Annu. Rev. Biophys. Biomol. Struct.*, **35**(1), 389–415, 2006.

[75] M. R. Shirts, J. W. Pitera, W. C. Swope, and V. S. Pande. Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins. *J. Chem. Phys.*, **119**(11), 5740–5761, 2003.

[76] M. R. Shirts and V. S. Pande. Solvation free energies of amino acid side chain analogs for common molecular mechanics water models. *J. Chem. Phys.*, **122**(13), 134508, 2005.

[77] D. Roe, A. Okur, L. Wickstrom, V. Hornak, and C. Simmerling. Secondary structure bias in generalized Born solvent models: comparison of conformational ensembles and free energy of solvent polarization from explicit and implicit solvation. *J. Phys. Chem. B*, **111**, 1846–1857, 2007.

[78] T. Raschke. Water structure and interactions with protein surfaces. *Curr. Opin. Struct. Biol.*, **16**, 152–159, 2006.

[79] V. Makarov, B. Pettitt, and M. Feig. Solvation and hydration of proteins and nucleic acids: a theoretical view of simulation and experiment. *Acc. Chem. Res.*, **35**, 376–384, 2002.

[80] C. Mattos. Protein-water interactions in a dynamic world. *Trends Biochem. Sci.*, **27**, 203–208, 2002.

[81] M. Weik. Low-temperature behavior of water confined by biological macromolecules and its relation to protein dynamics. *Eur. Phys. J. E Soft Matter*, **12**, 153–158, 2003.

[82] C. Lopez, R. Darst, and P. Rossky. Mechanistic Elements of Protein Cold Denaturation. *J. Phys. Chem. B*, 2008.

[83] A. Ben-Naim. Statistical mechanics of "waterlike" particles in two dimensions. i. Physical model and application of the Percus-Yevick equation. *J. Chem. Phys.*, **54**, 3682–3695, 1971.

[84] H. B. Yu and W. F. van Gunsteren. Charge-on-spring polarizable water models revisited: From water clusters to liquid water to ice. *J. Chem. Phys.*, **121**(19), 9549–9564, 2004.

[85] J. Lobaugh and G. A. Voth. A quantum model for water: Equilibrium and dynamical properties. *J. Chem. Phys.*, **106**(6), 2400–2410, 1997.

[86] M. I. Bernal-Uruchurtu, M. T. C. Martins-Costa, C. Millot, and M. F. Ruiz-López. Improving description of hydrogen bonds at the semiempirical level: water-water interactions as test case. *J. Comput. Chem.*, **21**(7), 572–581, 2000.

[87] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans. Interaction models for water in relation to protein hydration. *Intermolecular Forces*, pp. 331–342, 1981.

[88] C. Vega, E. Sanz, and J. L. F. Abascal. The melting temperature of the most common models of water. *J. Chem. Phys.*, **122**(11), 114507, 2005.

[89] W. L. Jorgensen. Quantum and statistical mechanical studies of liquids. 10. transferable intermolecular potential functions for water, alcohols, and ethers. application to liquid water. *J. Am. Chem. Soc.*, **103**(2), 335–340, 1981.

[90] J. D. Bernal and R. H. Fowler. A theory of water and ionic solution, with particular reference to hydrogen and hydroxyl ions. *J. Chem. Phys.*, **1**(8), 515–548, 1933.

[91] H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura, and T. Head-Gordon. Development of an improved four-site water model for biomolecular simulations: Tip4p-ew. *J. Chem. Phys.*, **120**(20), 9665–9678, 2004.

[92] M. W. Mahoney and W. L. Jorgensen. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J. Chem. Phys.*, **112**(20), 8910–8922, 2000.

[93] N. Baker. Improving implicit solvent simulations: a Poisson-centric view. *Curr. Opin. Struct. Biol.*, **15**, 137–143, 2005.

[94] M. Feig and C. Brooks. Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Curr. Opin. Struct. Biol.*, **14**, 217–224, 2004.

[95] N. Baker, D. Sept, S. Joseph, M. Holst, and J. McCammon. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.*, **21**, 10037–10041, 2001.

[96] T. Schlick, *Molecular modeling and simulation: An interdisciplinary Guide*, Spring-Verlag New York, Inc., 2002.

[97] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.*, **112**, 6127–6129, 1990.

[98] AMBER9 user's manual. D. A. Case, T. Darden, T. E. C. III, C. Simmerling, J. Wang, R. E. Duke, R. Luo, K. M. Merz, D. A. Pearlman, M. Crowley, R. Walker, WeiZhang, B. Wang, S. Hayik, A. Roitberg, G. Seabra, K. Wong, F. Paesani, X. Wu, S. Brozell, V. Tsui, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, P. Beroza, D. H. Mathews, C. Schafmeister, W. S. Ross, P. A. Kollman, R. V. Stanton, J. Pitera, I. Massova, A. Cheng, J. J. Vincent, R.'Radmer, G. L. Seibel, J. W.Caldwell, U. C. Singh, P. Weiner, P. Cieplak, Y. Duan, R. Woods, K. Kirschner, S. M. Tschampel, S. Weiner, A. Onufriev, C. Bayly, W. Cornell, and S. Weiner 2006.

[99] J. Srinivasan, M. W. Trevathan, P. Beroza, and D. A. Case. Application of a pairwise generalized Born model to proteins and nucleic acids: inclusion of salt effects. *Theoretical Chemistry Accounts*, **101**(6), 426–434, 1999.

[100] J. Chen and C. Brooks Iii. Implicit modeling of nonpolar solvation for simulating protein folding and conformational transitions. *Phys. Chem. Chem. Phys.*, **10**, 471–481, 2008.

[101] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, **4**, 187–217, 1983.

[102] A. MacKerel Jr., C. Brooks III, L. Nilsson, B. Roux, Y. Won, and M. Karplus, in *The Encyclopedia of Computational Chemistry*, ed. P. v. R. Schleyer et al., John Wiley & Sons: Chichester, 1998. Vol. 1, pp. 271–277.

[103] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen. GROMACS: fast, flexible, and free. *J. Comput. Chem.*, **26**(16), 1701–1718, 2005.

[104] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten. Scalable molecular dynamics with NAMD. *J. Comput. Chem.*, **26**(16), 1781–1802, 2005.

[105] L. Verlet. Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Phys. Rev.*, **159**(1), 98, 1967.

[106] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, Academic Press, 1996.

[107] M. S. Lee, J. Freddie R. Salsbury, and C. L. B. III. Novel generalized Born methods. *J. Chem. Phys.*, **116**(24), 10606–10614, 2002.

[108] A. Kovalenko and F. Hirata. Potential of mean force between two molecular ions in a polar molecular solvent: A study by the three-dimensional reference interaction site model. *J. Phys. Chem. B*, **103**(37), 7942–7957, 1999.

[109] W. van Gunsteren and H. Berendsen. Computer simulation of molecular dynamics: Methodology, applications and perspectives in chemistry. *Angew. Chemie. Int. Edit. Engl.*, **29**, 992–1023, 1990.

[110] R. Levy, M. Karplus, and J. McCammon. Diffusive langevin dynamics of model alkanes. *Chem. Phys. Lett.*, **65**, 4–11, 1979.

[111] A. Leach, *Molecular Modelling: Principles and Applications*, Addison Wesley Longman Ltd., 1996.

[112] S. Nosé. A molecular-dynamics method for simulations in the canonical ensemble. *Mol. Phys.*, **52**(2), 255–268, 1984.

[113] W. G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev., A Gen. Phys.*, **31**(3), 1695–1697, 1985.

[114] W. G. Hoover. Constant-pressure equations of motion. *Phys. Rev., A Gen. Phys.*, **34**(3), 2499–2500, 1986.

[115] H. Berendsen, J. Postma, W. van Gunsteren, A. DiNola, and J. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, **81**, 3684–3690, 1984.

[116] T. Morishita. Fluctuation formulas in molecular-dynamics simulations with the weak coupling heat bath. *J. Chem. Phys.*, **113**(8), 2976–2982, 2000.

[117] S. E. Feller, Y. Zhang, R. W. Pastor, and B. R. Brooks. Constant pressure molecular dynamics simulation: The langevin piston method. *J. Chem. Phys.*, **103**(11), 4613–4621, 1995.

[118] G. J. Martyna, D. J. Tobias, and M. L. Klein. Constant pressure molecular dynamics algorithms. *J. Chem. Phys.*, **101**(5), 4177–4189, 1994.

[119] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, Academic Press, second ed., 2002.

[120] A. D. MacKerell, Jr., D. Bashford, M. Bellott, R. Dunbrack Jr., J. Evanseck, M. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. Lau, C. Mattos, S. Michnick, T. Ngo, D. Nguyen, B. Prodhom, W. Reiher, III, B. Roux, M. Schlenkrich, J. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, **102**, 3586–3616, 1998.

[121] J. Wang, P. Cieplak, and P. A. Kollman. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.*, pp. 1049–1074, 2000.

[122] Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, and P. Kollman. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.*, **24**, 1999–2012, 2003.

[123] P. J. Steinbach and B. R. Brooks. New spherical-cutoff methods for long-range forces in macromolecular simulation. *J. Comput. Chem.*, **15**, 667–683, 1994.

[124] Chemistry at harvard macromolecular mechanics. President and Fellows of Harvard College.

[125] NAMD user's guide: Version 2.6. M. Bhandarkar, R. Brunner, C. Chipot, A. Dalke, S. Dixit, P. Grayson, J. Gullingsrud, A. Gursoy, D. Hardy, J. Hénin, W. Humphrey, D. Hurwitz, N. Krawetz, S. Kumar, M. Nelson, J. Phillips, A. Shinozaki, G. Zheng, and F. Zhu. The Board of Trustees of the University of Illinois, 2006.

[126] R. Garemyr and A. Elofsson. Study of the electrostatics treatment in molecular dynamics simulations. *Proteins: Struct. Func. Genet.*, **37**, 417–428, 1999.

[127] T. Darden, D. York, and L. Pedersen. Particle mesh ewald - an n.log(n) method for ewald sums in large systems. *J. Chem. Phys.*, **98**(12), 10089–10092, 1993.

[128] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen. A smooth particle mesh ewald method. *J. Chem. Phys.*, **103**(19), 8577–8593, 1995.

[129] C. Kittel, *Introduction to Solid State Physics*, John Wiley & Sons, sixth ed., 1986.

[130] S. Feller, R. Pastor, A. Rojnuckarin, S. Bogusz, and B. Brooks. Effect of electrostatic force truncation on interfacial and transport properties of water. *J. Phys. Chem.*, **100**, 17011–17020, 1996.

[131] B. A. Berg and T. Neuhaus. Multicanonical algorithms for 1st order phase-transitions. *Phys. Lett. B*, **267**(2), 249–253, 1991.

[132] B. A. Berg and T. Neuhaus. Multicanonical ensemble - a new approach to simulate 1st-order phase-transitions. *Phys. Rev. Lett.*, **68**(1), 9–12, 1992.

[133] A. Mitsutake, Y. Sugita, and Y. Okamoto. Generalized-ensemble algorithms for molecular simulations of biopolymers. *Biopolymers*, **60**(2), 96–123, 2001.

[134] U. Hansmann and Y. Okamoto. Prediction of peptide conformation by multicanonical algorithm: new approach to the multiple-minima problem. *J. Comput. Chem.*, **14**, 1333–1338, 1993.

[135] R. H. Swendsen and J.-S. Wang. Replica monte carlo simulation of spin-glasses. *Phys. Rev. Lett.*, **57**(21), 2607–2609, 1986.

[136] K. Hukushima and K. Nemoto. Exchange Monte Carlo method and application to spin glass simulations. *J Physical Soc Japan*, **65**(6), 1604–1608, 1996.

[137] M. C. Tesi, E. J. J. vanRensburg, E. Orlandini, and S. G. Whittington. Monte Carlo study of the interacting self-avoiding walk model in three dimensions. *J. Stat. Phys.*, **82**(1-2), 155–181, 1996.

[138] U. H. E. Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.*, **281**(1-3), 140–150, 1997.

[139] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, **314**(1-2), 141–151, 1999.

[140] T. Okabe, M. Kawata, Y. Okamoto, and M. Mikami. Replica-exchange monte carlo method for the isobaric-isothermal ensemble. *Chem. Phys. Lett.*, **335**, 435–439, 2001.

[141] C. Jarzynski. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.*, **78**(14), 2690–2693, 1997.

[142] S. Park, F. Khalili-Araghi, E. Tajkhorshid, and K. Schulten. Free energy calculation from steered molecular dynamics simulations using jarzynski's equality. *J. Chem. Phys.*, **119**(6), 3559–3566, 2003.

[143] E. Darve, M. A. Wilson, and A. Pohorille. Calculating free energies using a scaled-force molecular dynamics algorithm. *Mol. Simulat.*, **28**, 113–144, 2002.

[144] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **21**(6), 1087–1092, 1953.

[145] A. Mitsutake, M. Kinoshita, Y. Okamoto, and F. Hirata. Combination of the replica-exchange monte carlo method and the reference interaction site model theory for simulating a peptide molecule in aqueous solution. *J. Phys. Chem. B*, **108**(49), 19002–19012, 2004.

[146] GROMACS user manual version 3.3. D. van der Spoel, E. Lindahl, B. Hess, A. R. van Buuren, E. Apol, P. J. Meulenhoff, D. P. Tieleman, A. L. Sijbers, K. A. Feenstra, R. van Drunen, and H. J. Berendsen. www.gromacs.org, 2006.

[147] A. Okur, L. Wickstrom, M. Layten, R. Geney, K. Song, V. Hornak, and C. Simmerling. Improved efficiency of replica exchange simulations through use of a hybrid explicit/implicit solvation model. *J Chem Theory Comput*, **2**(2), 420–433, 2006.

[148] X. Cheng, G. Cui, V. Hornak, and C. Simmerling. Modified replica exchange simulation methods for local structure refinement. *J. Phys. Chem. B*, **109**, 8220–8230, 2005.

[149] Y. Rhee and V. Pande. Multiplexed-replica exchange molecular dynamics method for protein folding simulation. *Biophys. J.*, **84**, 775–786, 2003.

[150] M. J. RuizMontero, D. Frenkel, and J. J. Brey. Efficient schemes to compute diffusive barrier crossing rates. *Mol. Phys.*, **90**(6), 925–941, 1997.

[151] J. Hénin and C. Chipot. Overcoming free energy barriers using unconstrained molecular dynamics simulations. *J. Chem. Phys.*, **121**(7), 2904–2914, 2004.

[152] W. K. den Otter. Thermodynamic integration of the free energy along a reaction coordinate in cartesian coordinates. *J. Chem. Phys.*, **112**(17), 7283–7292, 2000.

[153] M. Sprik and G. Ciccotti. Free energy from constrained molecular dynamics. *J. Chem. Phys.*, **109**(18), 7737–7744, 1998.

[154] W. K. D. Otter and W. J. Briels. Free energy from molecular dynamics with multiple constraints. *Mol. Phys.*, **98**(12), 773–781, 2000.

[155] E. Darve and A. Pohorille. Calculating free energies using average force. *J. Chem. Phys.*, **115**(20), 9169–9183, 2001.

[156] D. Rodriguez-Gomez, E. Darve, and A. Pohorille. Assessing the efficiency of free energy calculation methods. *J. Chem. Phys.*, **120**(8), 3563–3578, 2004.

[157] T. P. Straatsma, H. Berendsen, and A. Stam. Estimation of statistical errors in molecular simulation calculations. *Mol. Phys.*, **57**(1), 89–95, 1986.

[158] G. Phillips and B. Pettitt. Structure and dynamics of the water around myoglobin. *Protein Sci.*, **4**, 149–158, 1995.

[159] T. Kamei, M. Oobatake, and M. Suzuki. Hydration of apomyoglobin in native, molten globule, and unfolded states by using microwave dielectric spectroscopy. *Biophys. J.*, **82**, 418–425, 2002.

[160] P. J. Steinbach and B. R. Brooks. Protein hydration elucidated by molecular dynamics simulation. *Proc. Natl. Acad. Sci. U.S.A.*, **90**, 9135–9139, 1993.

[161] P. J. Steinbach and B. R. Brooks. Hydrated myoglobin's anharmonic fluctuations are not primarily due to dihedral transitions. *Proc. Natl. Acad. Sci. U.S.A.*, **93**, 55–59, 1996.

[162] J. Smith, F. Merzel, A. Bondar, A. Tournier, and S. Fischer. Structure, dynamics and reactions of protein hydration water. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, **359**, 1181–1189, 2004.

[163] F. Merzel and J. Smith. Is the first hydration shell of lysozyme of higher density than bulk water? *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 5378–5383, 2002.

[164] J. KENDREW, G. BODO, H. DINTZIS, R. PARRISH, H. WYCKOFF, and D. PHILLIPS. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, **181**, 662–666, 1958.

[165] J. Kuriyan, S. Wilz, and G. A. Karplus, M.and Petsko. X-ray structure and refinement of carbon-monoxy (fe ii)-myoglobin at 1.5 a resolution. *J. Mol. Biol.*, **192**, 133, 1986. PDB ID: 1MBC.

[166] J. Ernst, R. Clubb, H. Zhou, A. Gronenborn, and G. Clore. Demonstration of positionally disordered water within a protein hydrophobic cavity by NMR. *Science*, **267**, 1813–1817, 1995.

[167] G. Clore, P. Wingfield, and A. Gronenborn. High-resolution three-dimensional structure of interleukin 1 beta in solution by three- and four-dimensional nuclear magnetic resonance spectroscopy. *Biochemistry*, **30**, 2315–2323, 1991.

[168] L. Zhang, L. Wang, Y. Kao, W. Qiu, Y. Yang, O. Okobiah, and D. Zhong. Mapping hydration dynamics around a protein surface. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 18461–18466, 2007.

[169] M. Tarek and D. Tobias. The role of protein-solvent hydrogen bond dynamics in the structural relaxation of a protein in glycerol versus water. *Eur. Biophys. J.*, 2008.

[170] M. Tarek and D. Tobias. The dynamics of protein hydration water: a quantitative comparison of molecular dynamics simulations and neutron-scattering experiments. *Biophys. J.*, **79**, 3244–3257, 2000.

[171] M. Tarek and D. J. Tobias. Effects of solvent damping on side chain and backbone contributions to the protein boson peak. *J. Chem. Phys.*, **115**(3), 1607–1612, 2001.

[172] R. H. Stote, D. J. States, and M. Karplus. On the treatment of electrostatic interactions in biomolecular simulation. *J. Chim. Phys*, **88**, 2149–2433, 1991.

[173] D. Sept, N. A. Baker, and J. A. McCammon. The physical basis of microtubule structure and stability. *Protein Sci.*, **12**(10), 2257–2261, 2003.

[174] P. Drabik, S. Gusarov, and A. Kovalenko. Microtubule stability studied by three-dimensional molecular theory of solvation. *Biophys. J.*, **92**, 394–403, 2007.

[175] E. M. Mandelkow, E. Mandelkow, and R. A. Milligan. Microtubule dynamics and microtubule caps: a time-resolved cryo-electron microscopy study. *J. Cell Biol.*, **114**(5), 977–991, 1991.

[176] B. Gigant, P. Curmi, C. Martin-Barbey, E. Charbaut, S. Lachkar, L. Lebeau, S. Siavoshian, A. Sobel, and M. Knossow. The 4 A X-ray structure of a tubulin:stathmin-like domain complex. *Cell*, **102**, 809–816, 2000.

[177] W. Hamilton. On a new species of imaginary quantities connected with a theory of quaternions. *Proc. R. Irish Acad.*, **2**, 424–434, 1844.

[178] W. Hamilton. On quaternions. *Proc. R. Irish Acad.*, **3**, 1–16, 1847.

[179] C. Karney. Quaternions in molecular modeling. *J. Mol. Graph. Model*, **25**, 595–604, 2007.

[180] B. K. P. Horn. Closed-form solution of absolute orientation using unit quaternions. *J. Opt. Soc. Am. A*, **4**(4), 629, 1987.

[181] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical recipes in C: the art of scientific computing*, Cambridge University Press, New York, NY, USA, 1988.

[182] A. Aksimentiev, I. Balabin, R. Fillingame, and K. Schulten. Insights into the molecular mechanism of rotation in the Fo sector of ATP synthase. *Biophys. J.*, **86**, 1332–1344, 2004.

[183] A. Fiser, R. K. Do, and A. Sali. Modeling of loops in protein structures. *Protein Sci.*, **9**(9), 1753–1773, 2000.

[184] H. Li, A. D. Robertson, and J. H. Jensen. Very fast empirical prediction and rationalization of protein pKa values. *Proteins: Struct. Func. Genet.*, **61**(4), 704–721, 2005.

[185] H. Xiao, P. Verdier-Pinard, N. Fernandez-Fuentes, B. Burd, R. Angeletti, A. Fiser, S. Horwitz, and G. Orr. Insights into the mechanism of microtubule stabilization by Taxol. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 10166–10173, 2006.

[186] F. Gittes, B. Mickey, J. Nettleton, and J. Howard. Flexural rigidity of microtubules and actin filaments measured from thermal fluctuations in shape. *J. Cell Biol.*, **120**(4), 923–934, 1993.

[187] B. Mickey and J. Howard. Rigidity of microtubules is increased by stabilizing agents. *J. Cell Biol.*, **130**, 909–917, 1995.

[188] R. Vale, C. Coppin, F. Malik, F. Kull, and R. Milligan. Tubulin GTP hydrolysis influences the structure, mechanical properties, and kinesin-driven transport of microtubules. *J. Biol. Chem.*, **269**, 23769–23775, 1994.

[189] H. Felgner, R. Frank, and M. Schliwa. Flexural rigidity of microtubules measured with the use of optical tweezers. *J. Cell. Sci.*, **109**, 509–516, 1996.

[190] K. Kawaguchi, S. Ishiwata, and T. Yamashita. Temperature dependence of the flexural rigidity of single microtubules. *Biochem. Biophys. Res. Commun.*, **366**, 637–642, 2008.

[191] E. Pai, U. Krengel, G. Petsko, R. Goody, W. Kabsch, and A. Wittinghofer. Refined crystal structure of the triphosphate conformation of H-ras p21 at 1.35 å resolution: implications for the mechanism of GTP hydrolysis. *EMBO J.*, **9**, 2351, 1990. PDB ID: 5P21.

[192] L. Tong, A. de Vos, M. Milburn, and S. Kim. Crystal structures at 2.2 å resolution of the catalytic domains of normal ras protein and an oncogenic mutant complexed with GDP. *J. Mol. Biol.*, **217**, 503, 1991.

[193] A. K. W. Leung, E. Lucile White, L. J. Ross, R. C. Reynolds, J. A. DeVito, and D. W. Borhani. Structure of Mycobacterium tuberculosis FtsZ Reveals Unexpected, G Protein-like Conformational Switches. *J. Mol. Biol.*, **342**(3), 953–970, 2004.

[194] E. D. Salmon. Microtubules: a ring for the depolymerization motor. *Curr. Biol.*, **15**(8), 299–302, 2005.

[195] E. Nogales, H. Wang, and H. Niederstrasser. Tubulin rings: which way do they curve? *Curr. Opin. Struct. Biol.*, **13**, 256–261, 2003.

[196] J. Hall, L. Dudley, P. Dobner, S. Lewis, and N. Cowan. Identification of two human beta-tubulin isotypes. *Mol. Cell Biol.*, **3**, 854–862, 1983.

[197] S. Lewis, M. Gilmartin, J. Hall, and N. Cowan. Three expressed sequences within the human beta-tubulin multigene family each define a distinct isotype. *J. Mol. Biol.*, **182**, 11–20, 1985.

[198] Q. Lu and R. Luduena. In vitro analysis of microtubule assembly of isotypically pure tubulin dimers. Intrinsic differences in the assembly properties of alpha beta II, alpha beta III, and alpha beta IV tubulin dimers in the absence of microtubule-associated proteins. *J. Biol. Chem.*, **269**, 2041–2047, 1994.

[199] K. Richards, K. Anders, E. Nogales, K. Schwartz, K. Downing, and D. Botstein. Structure-function relationships in yeast tubulins. *Mol. Biol. Cell*, **11**, 1887–1903, 2000.

[200] D. Panda, H. Miller, A. Banerjee, R. Luduea, and L. Wilson. Microtubule dynamics in vitro are regulated by the tubulin isotype composition. *Proc. Natl. Acad. Sci. U.S.A.*, **91**, 11358–11362, 1994.

[201] J. Fackenthal, F. Turner, and E. Raff. Tissue-specific microtubule functions in Drosophila spermatogenesis require the beta 2-tubulin isotype-specific carboxy terminus. *Dev. Biol.*, **158**, 213–227, 1993.

[202] J. Fackenthal, J. Hutchens, F. Turner, and E. Raff. Structural analysis of mutations in the Drosophila beta 2-tubulin isoform reveals regions in the beta-tubulin molecular required for general and for tissue-specific microtubule functions. *Genetics*, **139**, 267–286, 1995.

[203] V. Peyrot, C. Briand, and J. Andreu. C-terminal cleavage of tubulin by subtilisin enhances ring formation. *Arch. Biochem. Biophys.*, **279**, 328–337, 1990.

[204] B. Bhattacharyya, D. Sackett, and J. Wolff. Tubulin, hybrid dimers, and tubulin S. Stepwise charge reduction and polymerization. *J. Biol. Chem.*, **260**, 10208–10216, 1985.

[205] M. Mejillano, E. Tolo, R. Williams, and R. Himes. The conversion of tubulin carboxyl groups to amides has a stabilizing effect on microtubules. *Biochemistry*, **31**, 3478–3483, 1992.

[206] J. Wolff, D. Sackett, and L. Knipling. Cation selective promotion of tubulin polymerization by alkali metal chlorides. *Protein Sci.*, **5**, 2020–2028, 1996.

[207] S. Westermann, A. Avila-Sakar, H. Wang, H. Niederstrasser, J. Wong, D. Drubin, E. Nogales, and G. Barnes. Formation of a dynamic kinetochore- microtubule interface through assembly of the Dam1 ring complex. *Mol. Cell*, **17**, 277–290, 2005.

[208] L. Serrano, J. Avila, and R. Maccioni. Controlled proteolysis of tubulin by subtilisin: localization of the site for MAP2 interaction. *Biochemistry*, **23**, 4675–4681, 1984.

[209] V. Rodionov, F. Gyoeva, A. Kashina, S. Kuznetsov, and V. Gelfand. Microtubule-associated proteins and microtubule-based translocators have different binding sites on tubulin molecule. *J. Biol. Chem.*, **265**, 5702–5707, 1990.

[210] L. Knipling and J. Wolff. Direct interaction of Bcl-2 proteins with tubulin. *Biochem. Biophys. Res. Commun.*, **341**, 433–439, 2006.

[211] D. Rodi, R. Janes, H. Sanganee, R. Holton, B. Wallace, and L. Makowski. Screening of a library of phage-displayed peptides identifies human bcl-2 as a taxol-binding protein. *J. Mol. Biol.*, **285**, 197–203, 1999.

[212] M. Seibert, A. Patriksson, B. Hess, and D. van der Spoel. Reproducible polypeptide folding and structure prediction using molecular dynamics simulations. *J. Mol. Biol.*, **354**, 173–183, 2005.

[213] E. Nogales, S. Wolf, I. Khan, R. Luduea, and K. Downing. Structure of tubulin at 6.5 Å and location of the taxol-binding site. *Nature*, **375**, 424–427, 1995.

[214] W. L. Delano, The pymol molecular graphics system, 2002.

[215] E. Sorin and V. Pande. Exploring the helix-coil transition via all-atom equilibrium ensemble simulations. *Biophys. J.*, **88**, 2472–2493, 2005.

[216] E. Lindahl, B. Hess, and D. van der Spoel. Gromacs 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model.*, **7**(8), 306–317, 2001.

[217] X. Daura, K. Gademann, B. Jaun, D. Seebach, W. F. van Gunsteren, and A. E. Mark. Peptide folding: When simulation meets experiment. *Angew. Chemie. Int. Edit.*, **38**(1-2), 236–240, 1999.

[218] A. Garca. Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.*, **68**, 2696–2699, 1992.

[219] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen. Essential dynamics of proteins. *Proteins: Struct. Func. Genet.*, **17**(4), 412–425, 1993.

[220] V. Maiorov and G. Crippen. Size-independent comparison of protein three-dimensional structures. *Proteins: Struct. Func. Genet.*, **22**, 273–283, 1995.

[221] A. Amadei, B. de Groot, M. Ceruso, M. Paci, A. Di Nola, and H. Berendsen. A kinetic model for the internal motions of proteins: diffusion between multiple harmonic wells. *Proteins: Struct. Func. Genet.*, **35**, 283–292, 1999.

[222] B. Hess. Similarities between principal components of protein dynamics and random diffusion. *Phys. Rev., E Stat. Nonlin. Soft. Matter Phys.*, **62**, 8438–8448, 2000.

[223] B. Hess. Convergence of sampling in protein simulations. *Phys. Rev., E Stat. Nonlin. Soft. Matter Phys.*, **65**, 031910, 2002.

[224] D. Frishman and P. Argos. Knowledge-based protein secondary structure assignment. *Proteins: Struct. Func. Genet.*, **23**, 566–579, 1995.

[225] L. Falquet, M. Pagni, P. Bucher, N. Hulo, C. Sigrist, K. Hofmann, and A. Bairoch. The PROSITE database, its status in 2002. *Nucleic Acids Res.*, **30**, 235–238, 2002.

[226] B. M. Paschal, R. A. Obar, and R. B. Vallee. Interaction of brain cytoplasmic dynein and MAP2 with a common sequence at the C-terminus of tubulin. *Nature*, **342**(6249), 569–572, 1989.

[227] H. Grubmüller. Predicting slow structural transitions in macromolecular systems: Conformational flooding. *Phys. Rev., E Stat. Nonlin. Soft. Matter Phys.*, **52**, 2893–2906, 1995.

[228] I. Kosztin, B. Barz, and L. Janosi. Calculating potentials of mean force and diffusion coefficients from nonequilibrium processes without Jarzynski's equality. *J. Chem. Phys.*, **124**(6), 2006.

[229] B. Ma, C. Tsai, and R. Nussinov. A systematic study of the vibrational free energies of polypeptides in folded and random states. *Biophys. J.*, **79**, 2739–2753, 2000.

[230] B. Hess. Determining the shear viscosity of model liquids from molecular dynamics simulations. *J. Chem. Phys.*, **116**(1), 209–217, 2002.

[231] K. Sugase, H. Dyson, and P. Wright. Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature*, **447**, 1021–1025, 2007.

[232] M. Plischke and B. Bergersen, *Equilibrium Statistical Physics*, World Scientific, 2nd ed., 1994.

[233] N. Cowan and L. Dudley. Tubulin isotypes and the multigene tubulin families. *Int. Rev. Cytol.*, **85**, 147–173, 1983.

[234] H. Jensen-Smith, R. Ludueña, and R. Hallworth. Requirement for the betaI and betaIV tubulin isotypes in mammalian cilia. *Cell Motil. Cytoskeleton*, **55**, 213–220, 2003.

[235] K. Sullivan and D. Cleveland. Identification of conserved isotype-defining variable region sequences for four vertebrate beta tubulin polypeptide classes. *Proc. Natl. Acad. Sci. U.S.A.*, **83**, 4327–4331, 1986.

[236] S. Ranganathan, D. Dexter, C. Benetatos, A. Chapman, K. Tew, and G. Hudes. Increase of beta(III)- and beta(IVa)-tubulin isotopes in human prostate carcinoma cells as a result of estramustine resistance. *Cancer Res.*, **56**, 2584–2589, 1996.

[237] M. Kavallaris, C. Burkhart, and S. Horwitz. Antisense oligonucleotides to class III beta-tubulin sensitize drug-resistant cells to Taxol. *Br. J. Cancer*, **80**, 1020–1025, 1999.

[238] S. Mozzetti, C. Ferlini, P. Concolino, F. Filippetti, G. Raspaglio, S. Prislei, D. Gallo, E. Martinelli, F. Ranelletti, G. Ferrandina, and G. Scambia. Class III beta-tubulin overexpression is a prominent mechanism of paclitaxel resistance in ovarian cancer patients. *Clin. Cancer Res.*, **11**, 298–305, 2005.

[239] M. Kavallaris, A. Tait, B. Walsh, L. He, S. Horwitz, M. Norris, and M. Haber. Multiple microtubule alterations are associated with Vinca alkaloid resistance in human leukemia cells. *Cancer Res.*, **61**, 5803–5809, 2001.

[240] D. Chandler and H. C. Andersen. Optimized cluster expansions for classical fluids. ii. theory of molecular liquids. *J. Chem. Phys.*, **57**(5), 1930–1937, 1972.

[241] F. Hirata and P. J. Rossky. An extended RISM equation for molecular polar fluids. *Chem. Phys. Lett.*, pp. 329–334, 1981.

[242] F. Hirata, B. M. Pettitt, and P. J. Rossky. Application of an extended rism equation to dipolar and quadrupolar fluids. *J. Chem. Phys.*, **77**(1), 509–520, 1982.

[243] F. Hirata, P. J. Rossky, and B. M. Pettitt. The interionic potential of mean force in a molecular polar solvent from an extended rism equation. *J. Chem. Phys.*, **78**(6), 4133–4144, 1983.

[244] T. Miyata and F. Hirata. Combination of molecular dynamics method and 3D-RISM theory for conformational sampling of large flexible molecules in solution. *J. Comput. Chem.*, **29**, 871–882, 2007.

[245] M. Kinoshita In Hirata [251], chapter 3.

[246] K. Schmidt and S. Kast. Hybrid integral equation/monte carlo approach to complexation thermodynamics. *J. Phys. Chem. B*, **106**(24), 6289–6297, 2002.

[247] M. Kinoshita, Y. Okamoto, and F. Hirata. First-principle determination of peptide conformations in solvents: Combination of monte carlo simulated annealing and rism theory. *J. Am. Chem. Soc.*, **120**(8), 1855–1863, 1998.

[248] S. Gusarov, T. Ziegler, and A. Kovalenko. Self-consistent combination of the three-dimensional rism theory of molecular solvation with analytical gradients and the amsterdam density functional package. *J. Phys. Chem. A*, **110**(18), 6083–6090, 2006.

[249] F. Hirata In *Molecular Theory of Solvation* [251], chapter 1.

[250] A. Kovalenko In Hirata [251], chapter 4.

[251] ed. F. Hirata, *Molecular Theory of Solvation*, Kluwer Academic Publishers, 2003.

[252] S. J. Singer and D. Chandler. Free-energy functions in the extended rism approximation. *Mol. Phys.*, **55**(3), 621–625, 1985.

[253] J. Burkardt, BLEND: Transfinite interpolation.

[254] A. Kovalenko, S. Ten-No, and F. Hirata. Solution of three-dimensional reference interaction site model and hypernetted chain equations for simple point charge water by modified method of direct inversion in iterative subspace. *J. Comput. Chem.*, **20**(9), 928–936, 1999.

[255] M. Frigo and S. G. Johnson. The design and implementation of FFTW3. *Proc. IEEE*, **93**(2), 216–231, 2005. special issue on "Program Generation, Optimization, and Platform Adaptation".

[256] R. J. Loncharich, B. R. Brooks, and R. W. Pastor. Langevin dynamics of peptides: the frictional dependence of isomerization rates of N-acetylalanyl-N'-methylamide. *Biopolymers*, **32**(5), 523–535, 1992.

[257] B. M. Pettitt and P. J. Rossky. Integral-equation predictions of liquid-state structure for waterlike intermolecular potentials. *J. Chem. Phys.*, **77**(3), 1451–1457, 1982.

[258] B. M. Pettitt and P. J. Rossky. The contribution of hydrogen-bonding to the structure of liquid methanol. *J. Chem. Phys.*, **78**(12), 7296–7299, 1983.

[259] F. Hirata and R. M. Levy. A new rism integral-equation for solvated polymers. *Chem. Phys. Lett.*, **136**(3-4), 267–273, 1987.

[260] F. Hirata and R. M. Levy. Ionic association in methanol and related solvents - an extended rism analysis. *J. Phys. Chem.*, **91**(18), 4788–4795, 1987.

[261] A. Kovalenko and F. Hirata. Potentials of mean force of simple ions in ambient aqueous solution. i. three-dimensional reference interaction site model approach. *J. Chem. Phys.*, **112**(23), 10391–10402, 2000.

[262] A. Cochran, N. Skelton, and M. Starovasnik. Tryptophan zippers: stable, monomeric beta -hairpins. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 5578–5583, 2001.

[263] R. Johnson, T. Yamazaki, A. Kovalenko, and H. Fenniri. Molecular basis for water-promoted supramolecular chirality inversion in helical rosette nanotubes. *J. Am. Chem. Soc.*, **129**(17), 5735–5743, 2007.

[264] T. Yamazaki, T. Imai, F. Hirata, and A. Kovalenko. Theoretical study of the cosolvent effect on the partial molar volume change of staphylococcal nuclease associated with pressure denaturation. *J. Phys. Chem. B*, **111**(5), 1206–1212, 2007.

[265] T. Imai, S. Ohyama, A. Kovalenko, and F. Hirata. Theoretical study of the partial molar volume change associated with the pressure-induced structural transition of ubiquitin. *Protein Sci.*, **16**, 1927–1933, 2007.

[266] T. Imai, R. Hiraoka, T. Seto, A. Kovalenko, and F. Hirata. Three-dimensional distribution function theory for the prediction of protein-ligand binding sites and affinities: application to the binding of noble gases to hen egg-white lysozyme in aqueous solution. *J. Phys. Chem. B*, **111**, 11585–11591, 2007.

[267] T. Imai, R. Hiraoka, A. Kovalenko, and F. Hirata. Water molecules in a protein cavity detected by a statistical-mechanical theory. *J. Am. Chem. Soc.*, **127**(44), 15334–15335, 2005.

[268] T. Imai, R. Hiraoka, A. Kovalenko, and F. Hirata. Locating missing water molecules in protein cavities by the three-dimensional reference interaction site model theory of molecular solvation. *Proteins: Struct. Func. Genet.*, **66**, 804–813, 2007.

[269] J. Perkyns and B. M. Pettitt. A site–site theory for finite concentration saline solutions. *J. Chem. Phys.*, **97**(10), 7656–7666, 1992.

[270] J. S. Perkyns and B. M. Pettitt. A dielectrically consistent interaction site theory for solvent electrolyte mixtures. *Chem. Phys. Lett.*, **190**(6), 626–630, 1992.

[271] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Struct. Func. Genet.*, **65**, 712–725, 2006.

[272] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical-integration of cartesian equations of motion of a system with constraints - molecular-dynamics of n-alkanes. *J. Comput. Phys.*, **23**(3), 327–341, 1977.

[273] J. Mongan, C. Simmerling, J. McCammon, D. Case, and A. Onufriev. Generalized born model with a simple, robust molecular volume correction. *J Chem Theory Comput*, **3**(1), 156–169, 2007.

[274] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in Fortran*, Cambridge University Press, 1992.

[275] H. Wang, V. Ramey, S. Westermann, A. Leschziner, J. Welburn, Y. Nakajima, D. Drubin, G. Barnes, and E. Nogales. Architecture of the Dam1 kinetochore ring complex and implications for microtubule-driven assembly and force-coupling mechanisms. *Nat. Struct. Mol. Biol.*, **14**, 721–726, 2007.

[276] E. Grishchuk, M. Molodtsov, F. Ataullakhanov, and J. McIntosh. Force production by disassembling microtubules. *Nature*, **438**, 384–388, 2005.

[277] M. I. Molodtsov, E. L. Grishchuk, A. K. Efremov, J. R. McIntosh, and F. I. Ataullakhanov. Force production by depolymerizing microtubules: a theoretical study. *Proc. Natl. Acad. Sci. U.S.A.*, **102**(12), 4353–4358, 2005.

[278] L. Amos and J. Löwe. How Taxol® stabilises micrtotubule structure. *Chem. Biol.*, **6**, R65–R69, 1999.

[279] A. L. Lehninger, *Principles of Biochemistry*, Worth Publishers, Inc., 1982.

[280] H. Li, A. D. Robertson, and J. H. Jensen. The determinants of carboxyl pKa values in turkey ovomucoid third domain. *Proteins: Struct. Func. Genet.*, **55**(3), 689–704, 2004. Evaluation Studies.

[281] J. Mongan, D. A. Case, and J. A. McCammon. Constant pH molecular dynamics in generalized Born implicit solvent. *J. Comput. Chem.*, **25**(16), 2038–2048, 2004.

[282] S. S. Zummdahl, *Chemistry*, D. C. Heath and Company, 1993.

[283] S. Elliott, *The Physics and Chemistry of Solids*, John Wiley & Sons Ltd, 2000.

[284] ed. A. T. Fojo, *Microtubule Targets in Cancer Therapy*, Humana Press, 2007.